Applied genomics

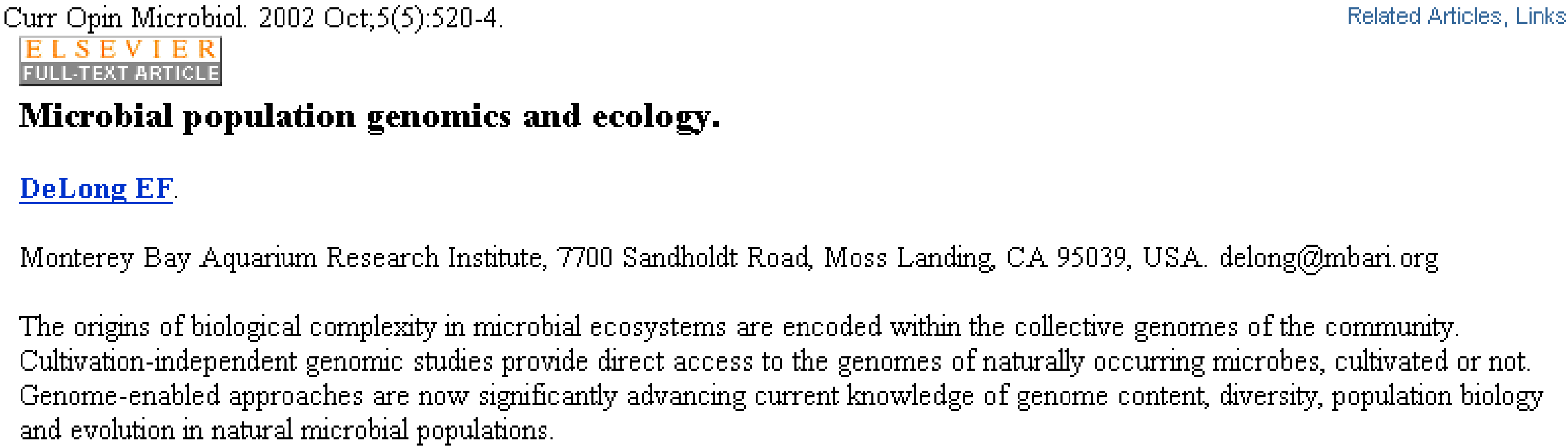
Metagenomics

Prof. Alberto Pallavicini
pallavic@units.it

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- **Metagenomics** is 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

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



Curr Opin Microbiol. 2002 Oct;5(5):520-4.

Related Articles, Links

ELSEVIER
FULL-TEXT ARTICLE

Microbial population genomics and ecology.

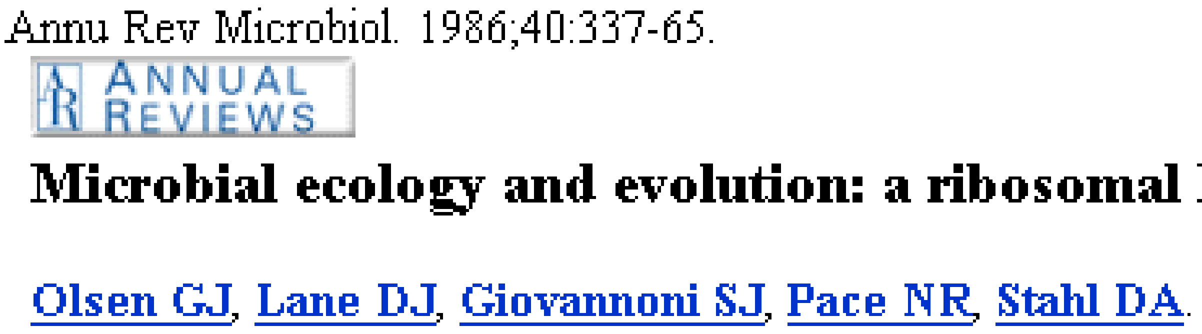
DeLong EF.

Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, CA 95039, USA. delong@mbari.org

The origins of biological complexity in microbial ecosystems are encoded within the collective genomes of the community. Cultivation-independent genomic studies provide direct access to the genomes of naturally occurring microbes, cultivated or not. Genome-enabled approaches are now significantly advancing current knowledge of genome content, diversity, population biology and evolution in natural microbial populations.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- The field has its roots in the culture-independent retrieval of genes, pioneered by Pace and colleagues two decades ago



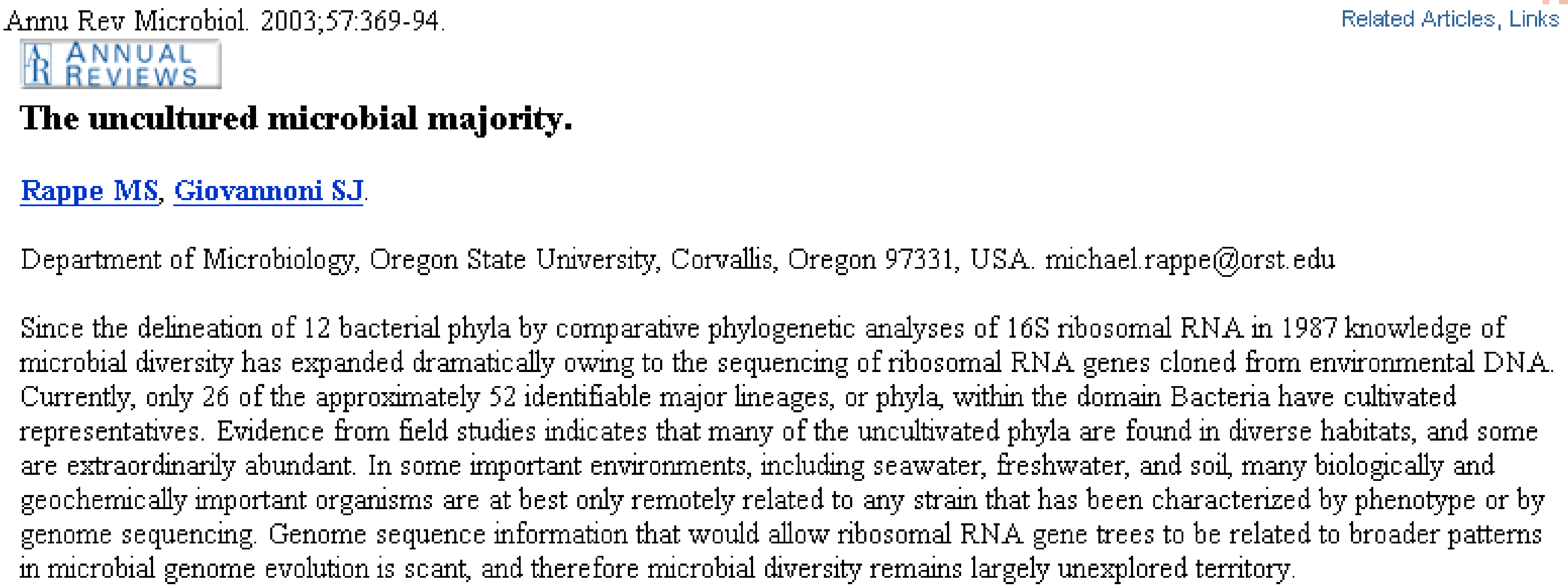
Annu Rev Microbiol. 1986;40:337-65.

Microbial ecology and evolution: a ribosomal RNA approach.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- Since then, metagenomics has revolutionized microbiology by shifting focus away from clonal isolates towards the estimated 99% of microbial species that cannot currently be cultivated.



Annu Rev Microbiol. 2003;57:369-94.

Related Articles, Links

The uncultured microbial majority.

Rappe MS, Giovannoni SJ.

Department of Microbiology, Oregon State University, Corvallis, Oregon 97331, USA. michael.rappe@orst.edu

Since the delineation of 12 bacterial phyla by comparative phylogenetic analyses of 16S ribosomal RNA in 1987 knowledge of microbial diversity has expanded dramatically owing to the sequencing of ribosomal RNA genes cloned from environmental DNA. Currently, only 26 of the approximately 52 identifiable major lineages, or phyla, within the domain Bacteria have cultivated representatives. Evidence from field studies indicates that many of the uncultivated phyla are found in diverse habitats, and some are extraordinarily abundant. In some important environments, including seawater, freshwater, and soil, many biologically and geochemically important organisms are at best only remotely related to any strain that has been characterized by phenotype or by genome sequencing. Genome sequence information that would allow ribosomal RNA gene trees to be related to broader patterns in microbial genome evolution is scant, and therefore microbial diversity remains largely unexplored territory.

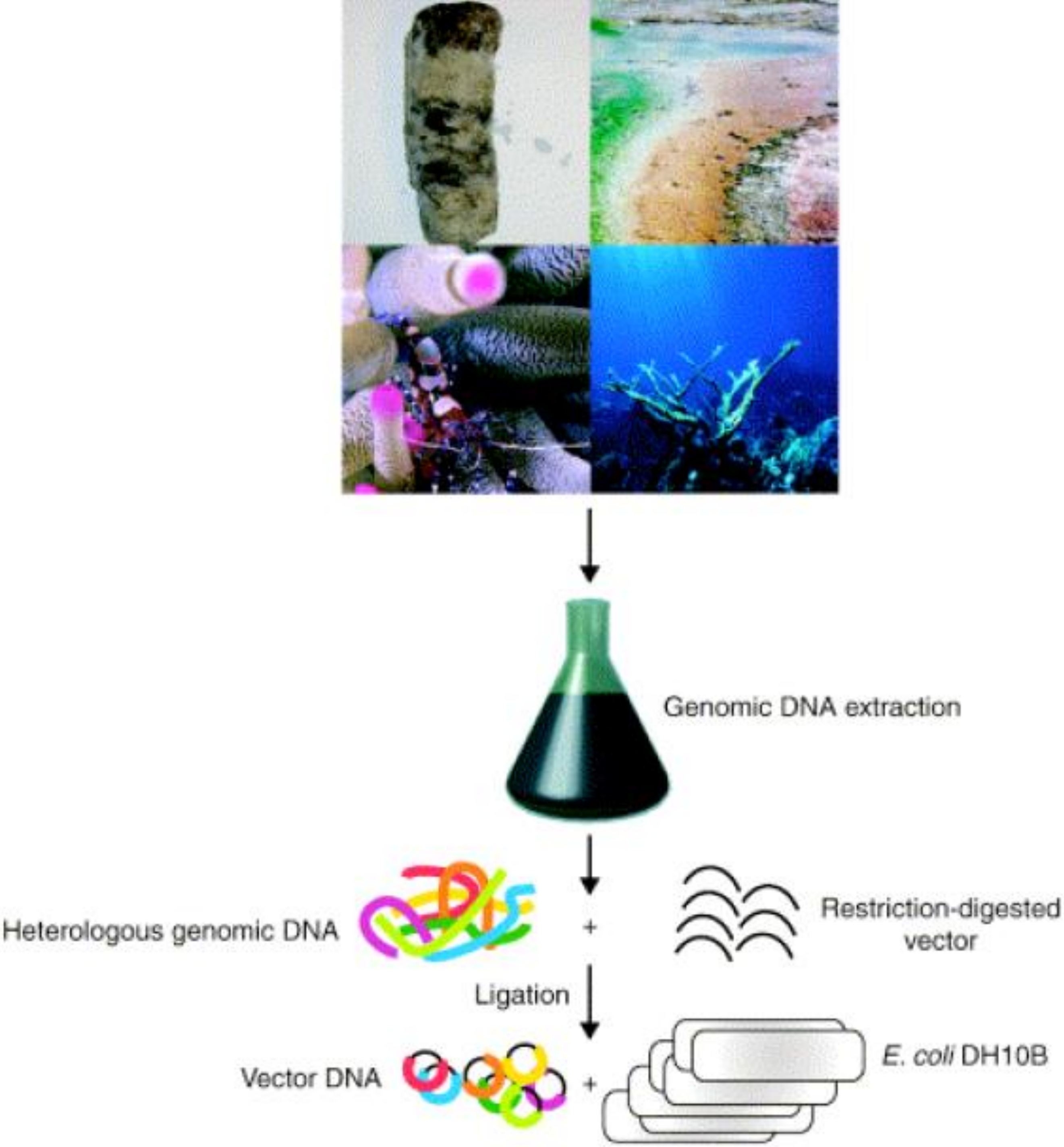
Metagenomics

- Metagenomics for biotechnological purposes
- Metagenomics for biomedical purposes
- Metagenomics for ecological analysis

- Whole genome metagenomics
- Gene centric metagenomics

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- At the beginnin a typical metagenomics project begins with the construction of a clone library from DNA sequence retrieved from an environmental sample.
- Clones are then selected for sequencing using either functional or sequence-based screens.



Genomic DNA extraction

Heterologous genomic DNA

+

Restriction-digested vector

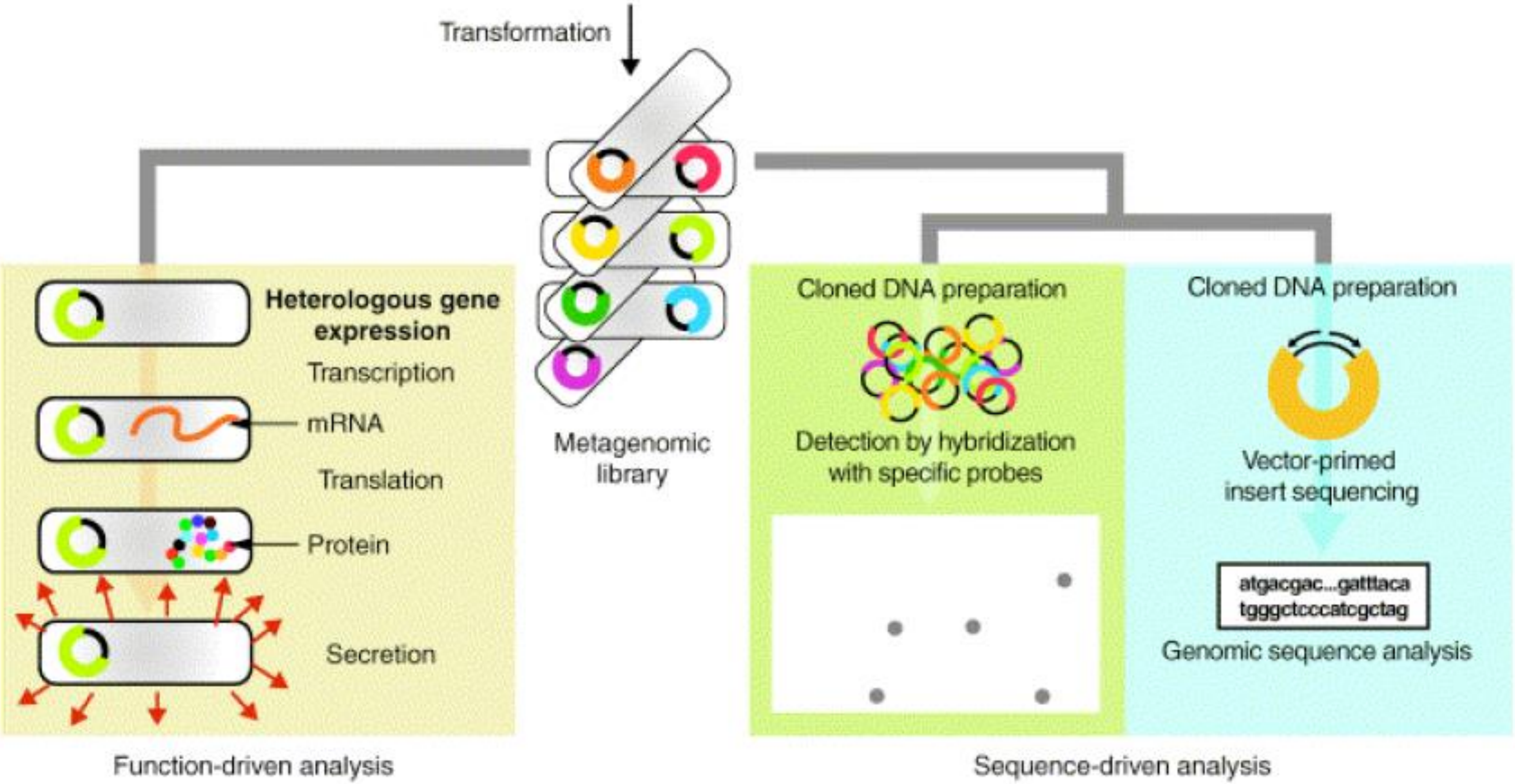
Ligation

Vector DNA

+

E. coli DH10B

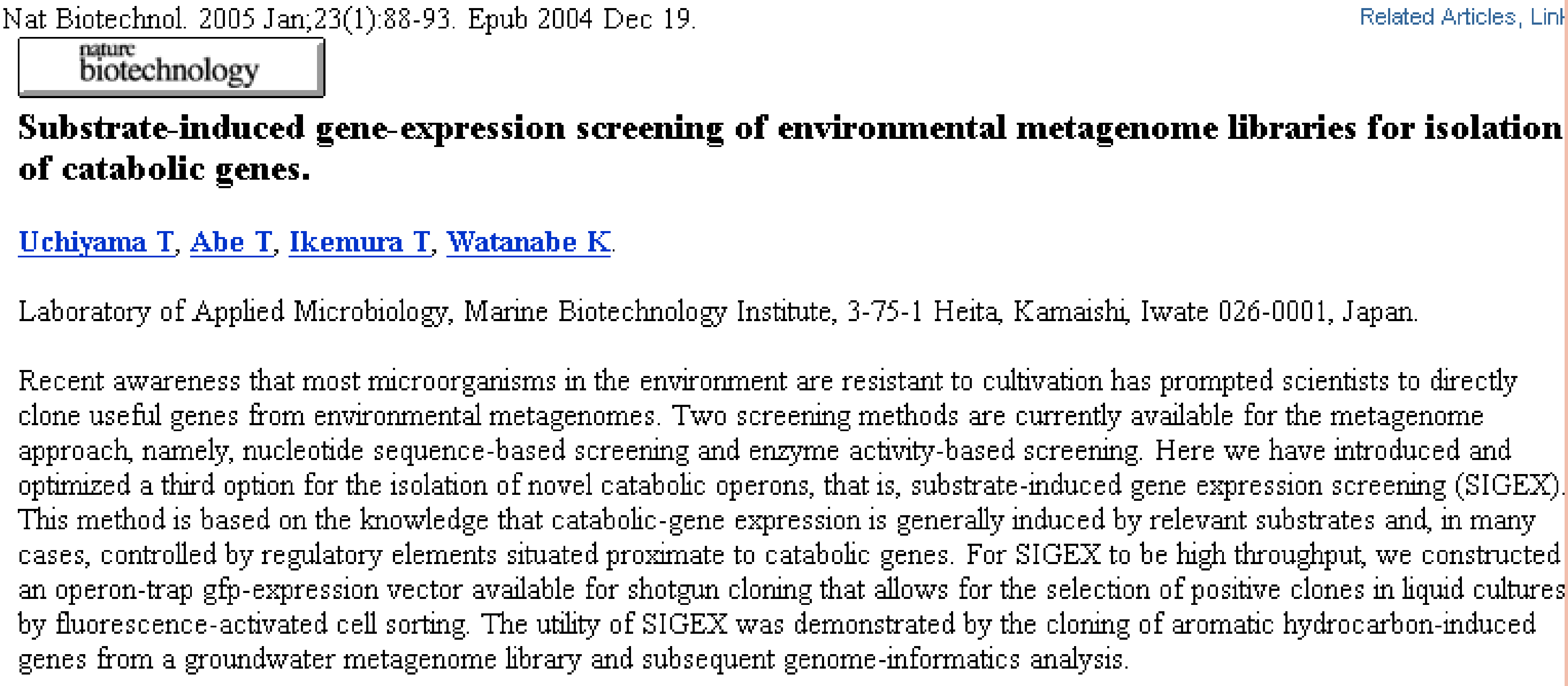
Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

In the **functional approach**, genes retrieved from the environment are heterologously expressed in a host, such as *Escherichia coli*, and sophisticated functional screens employed to detect clones expressing functions of interest.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



Nat Biotechnol. 2005 Jan;23(1):88-93. Epub 2004 Dec 19.

Related Articles, Link

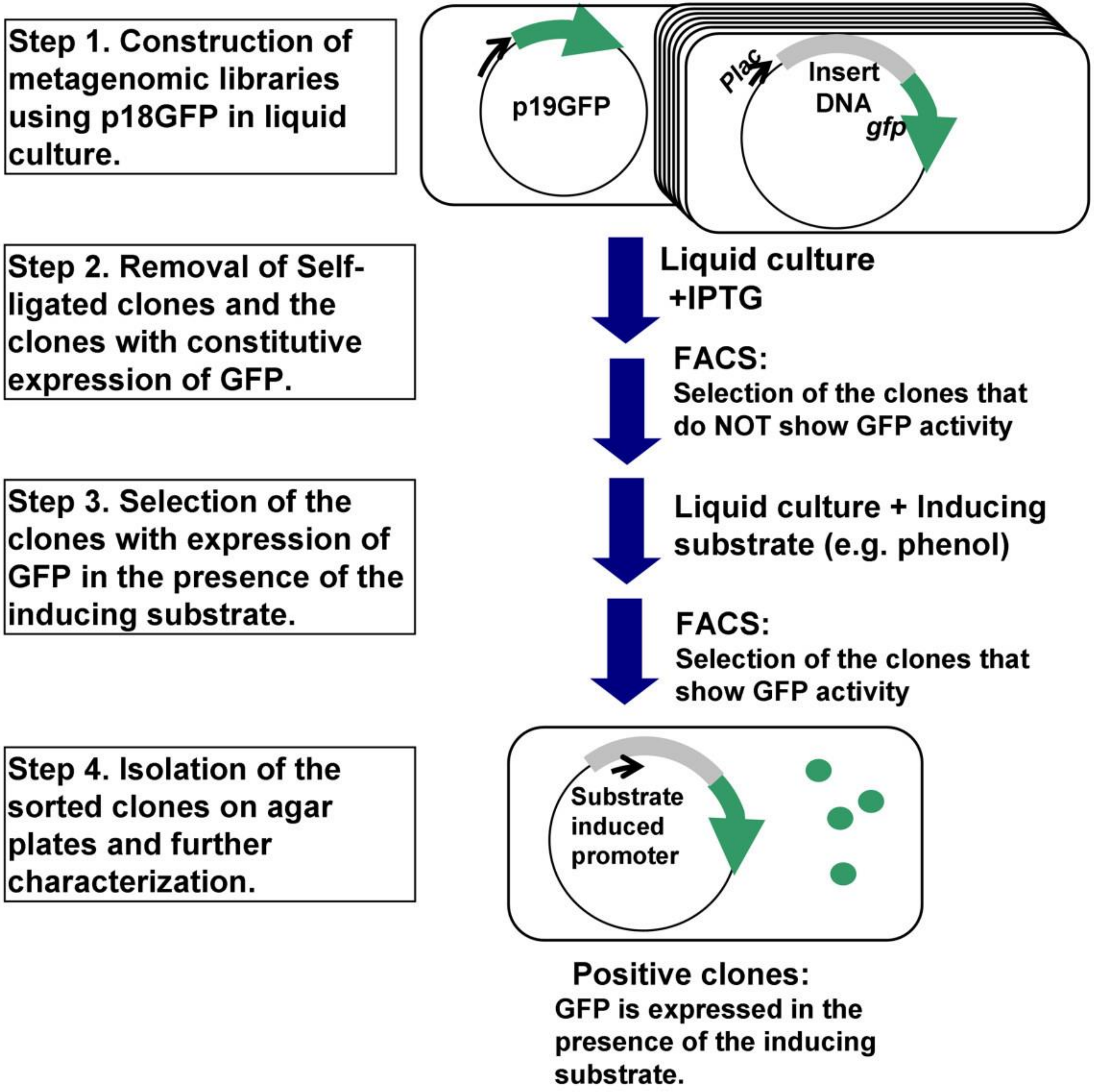
Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes.

Uchiyama T, Abe T, Ikemura T, Watanabe K.

Laboratory of Applied Microbiology, Marine Biotechnology Institute, 3-75-1 Heita, Kamaishi, Iwate 026-0001, Japan.

Recent awareness that most microorganisms in the environment are resistant to cultivation has prompted scientists to directly clone useful genes from environmental metagenomes. Two screening methods are currently available for the metagenome approach, namely, nucleotide sequence-based screening and enzyme activity-based screening. Here we have introduced and optimized a third option for the isolation of novel catabolic operons, that is, substrate-induced gene expression screening (SIGEX). This method is based on the knowledge that catabolic-gene expression is generally induced by relevant substrates and, in many cases, controlled by regulatory elements situated proximate to catabolic genes. For SIGEX to be high throughput, we constructed an operon-trap gfp-expression vector available for shotgun cloning that allows for the selection of positive clones in liquid cultures by fluorescence-activated cell sorting. The utility of SIGEX was demonstrated by the cloning of aromatic hydrocarbon-induced genes from a groundwater metagenome library and subsequent genome-informatics analysis.

To design of SIGEX is based on the facts that the expression of catabolic genes is generally **induced by substrates or metabolites of catabolic enzymes**, and that the expression of catabolic genes is controlled by regulatory elements located proximately in many cases.



Step 1. Construction of metagenomic libraries using p18GFP in liquid culture.

Step 2. Removal of Self-ligated clones and the clones with constitutive expression of GFP.

Step 3. Selection of the clones with expression of GFP in the presence of the inducing substrate.

Step 4. Isolation of the sorted clones on agar plates and further characterization.

Liquid culture +IPTG

FACS:
Selection of the clones that do NOT show GFP activity

Liquid culture + Inducing substrate (e.g. phenol)

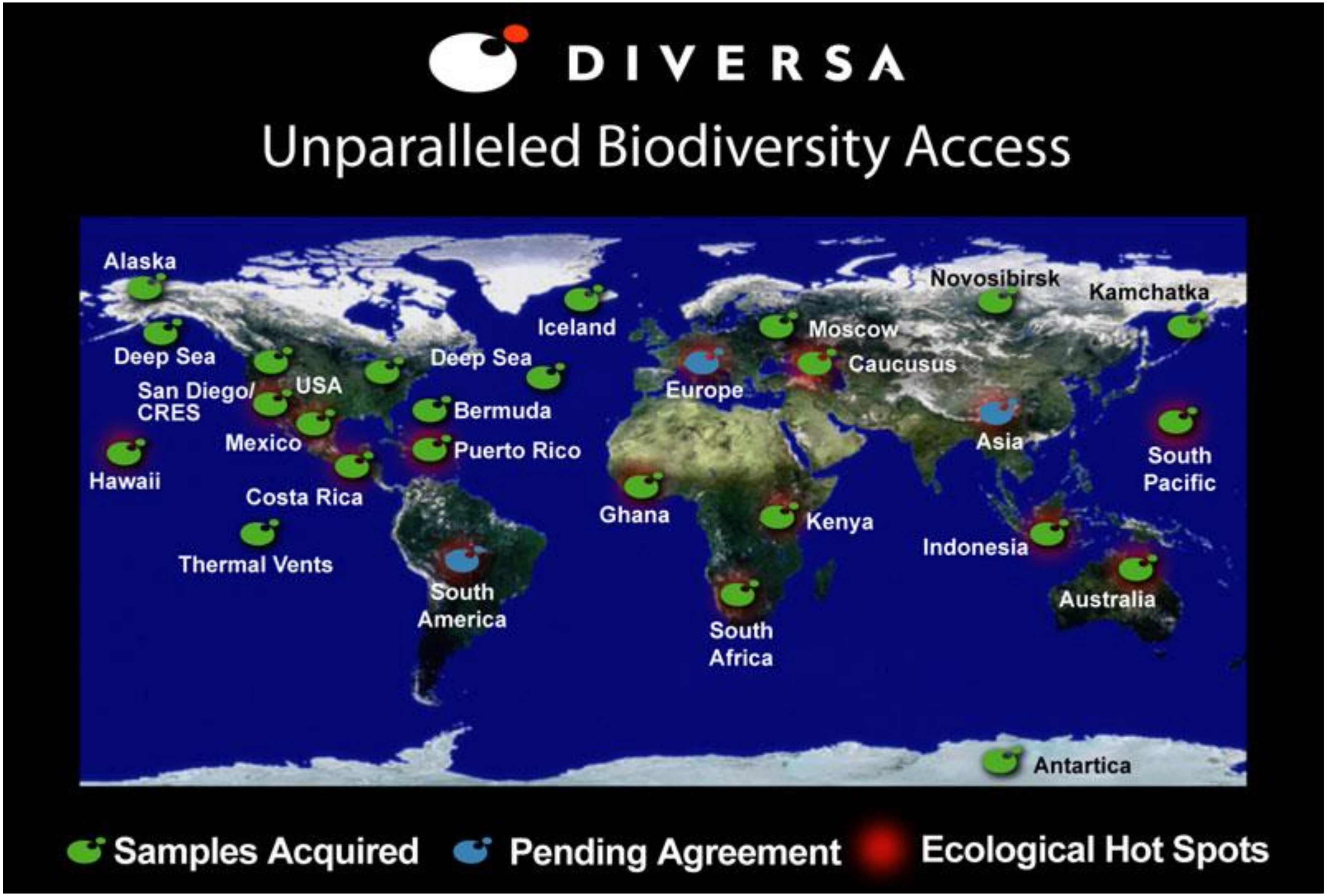
FACS:
Selection of the clones that show GFP activity

Positive clones:
GFP is expressed in the presence of the inducing substrate.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- This approach has produced many exciting discoveries and spawned several companies aiming to retrieve marketable natural products from the environment.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



DIVERSA

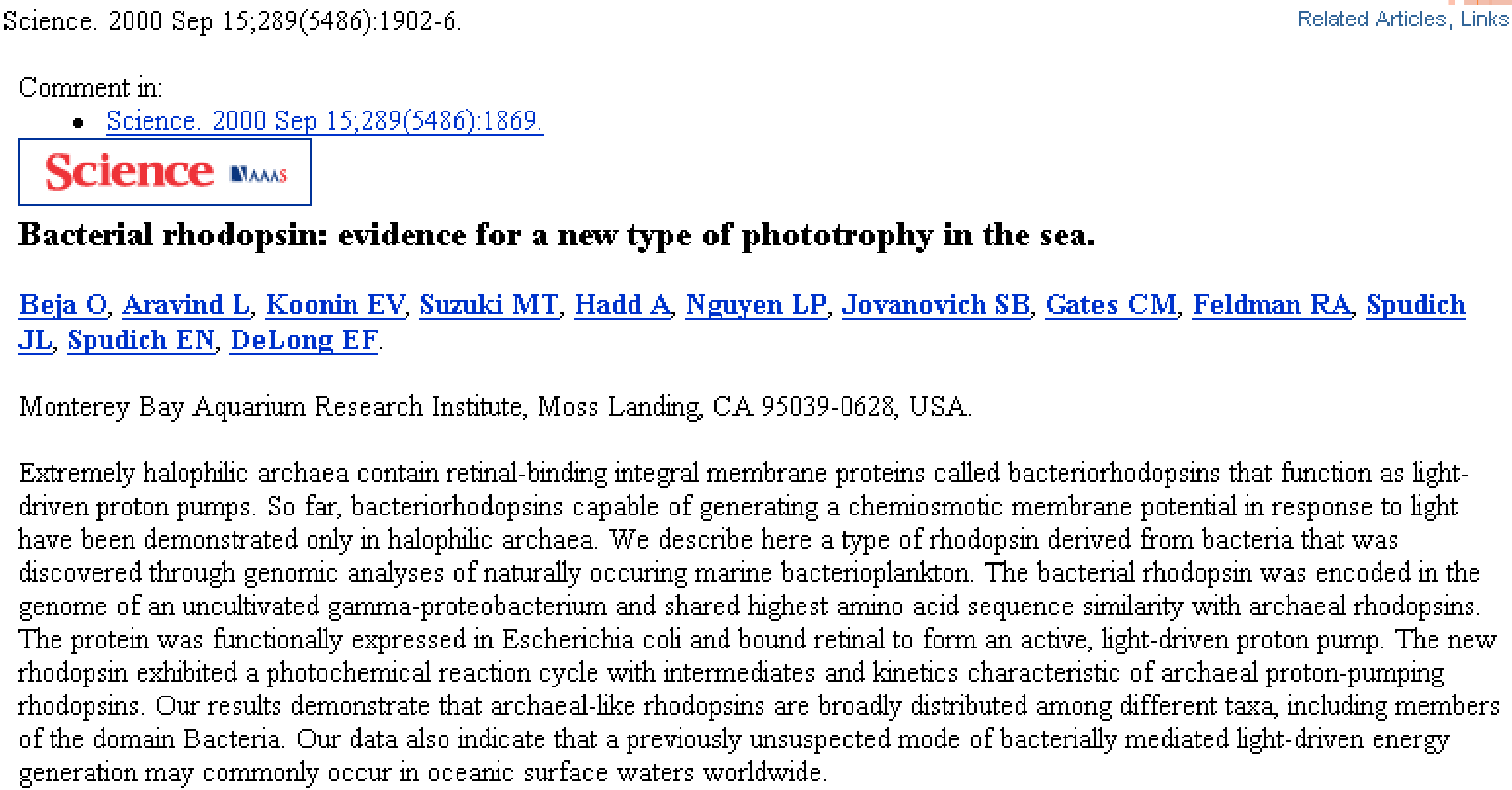
Unparalleled Biodiversity Access

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- In the sequence-based approach, clones are selected for sequencing based on the presence of genes of biological interest.

- One of the first discovery from this approach thus far is the **proteorhodopsin** gene from a marine community

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



Science. 2000 Sep 15;289(5486):1902-6.

Related Articles, Links

Comment in:

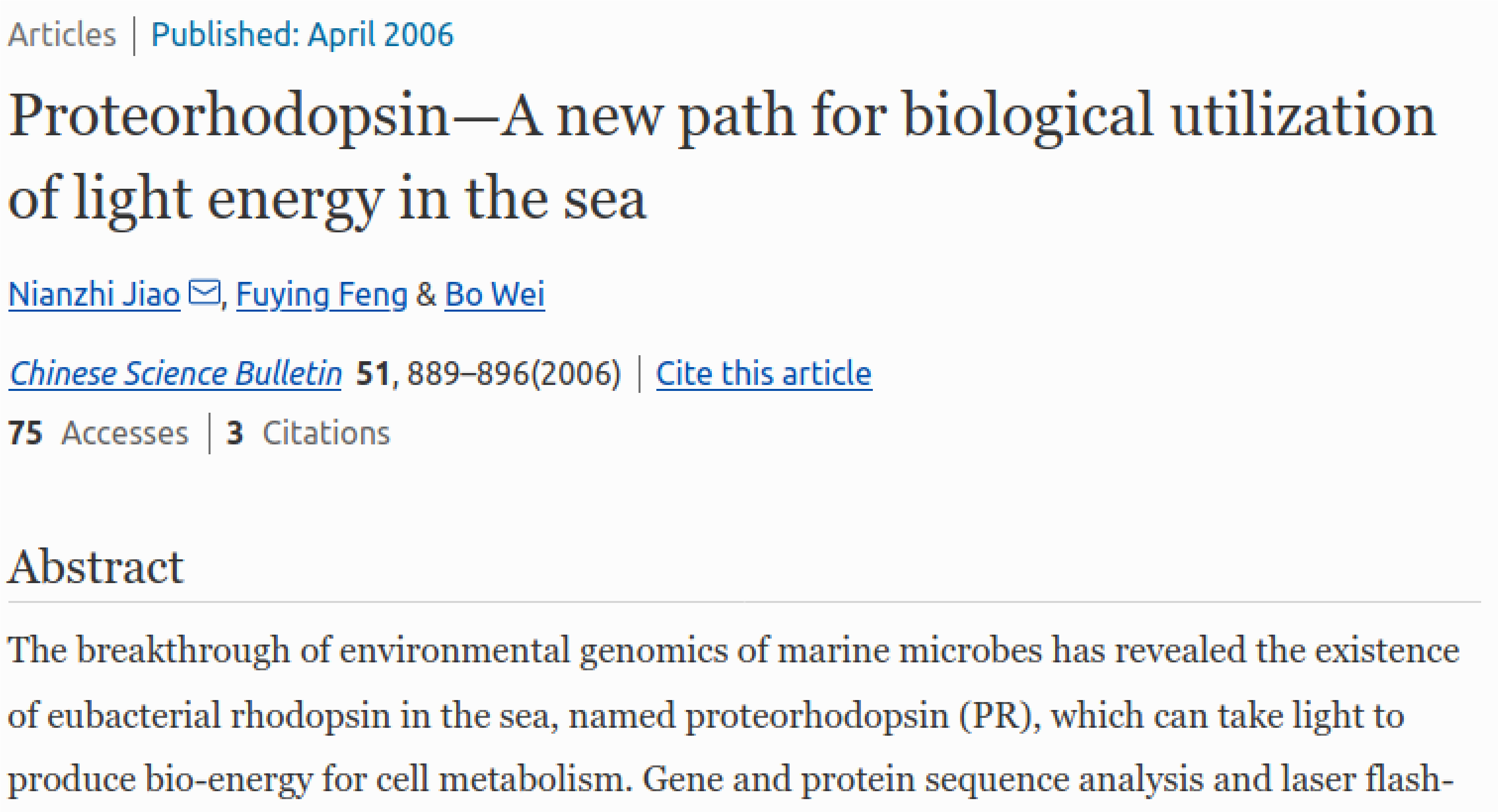
- Science. 2000 Sep 15;289(5486):1869.

Bacterial rhodopsin: evidence for a new type of phototrophy in the sea.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF.

Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039-0628, USA.

Extremely halophilic archaea contain retinal-binding integral membrane proteins called bacteriorhodopsins that function as light-driven proton pumps. So far, bacteriorhodopsins capable of generating a chemiosmotic membrane potential in response to light have been demonstrated only in halophilic archaea. We describe here a type of rhodopsin derived from bacteria that was discovered through genomic analyses of naturally occuring marine bacterioplankton. The bacterial rhodopsin was encoded in the genome of an uncultivated gamma-proteobacterium and shared highest amino acid sequence similarity with archaeal rhodopsins. The protein was functionally expressed in Escherichia coli and bound retinal to form an active, light-driven proton pump. The new rhodopsin exhibited a photochemical reaction cycle with intermediates and kinetics characteristic of archaeal proton-pumping rhodopsins. Our results demonstrate that archaeal-like rhodopsins are broadly distributed among different taxa, including members of the domain Bacteria. Our data also indicate that a previously unsuspected mode of bacterially mediated light-driven energy generation may commonly occur in oceanic surface waters worldwide.



Articles | Published: April 2006

Proteorhodopsin—A new path for biological utilization of light energy in the sea

Nianzhi Jiao ✉, Fuying Feng & Bo Wei

Chinese Science Bulletin **51**, 889–896(2006) | Cite this article

75 Accesses | **3** Citations

Abstract

The breakthrough of environmental genomics of marine microbes has revealed the existence of eubacterial rhodopsin in the sea, named proteorhodopsin (PR), which can take light to produce bio-energy for cell metabolism. Gene and protein sequence analysis and laser flash-

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- Recently, facilitated by the increasing capacity of sequencing centers, whole-genome shotgun (WGS) sequencing of the entire clone library has emerged as a third approach to metagenomics.

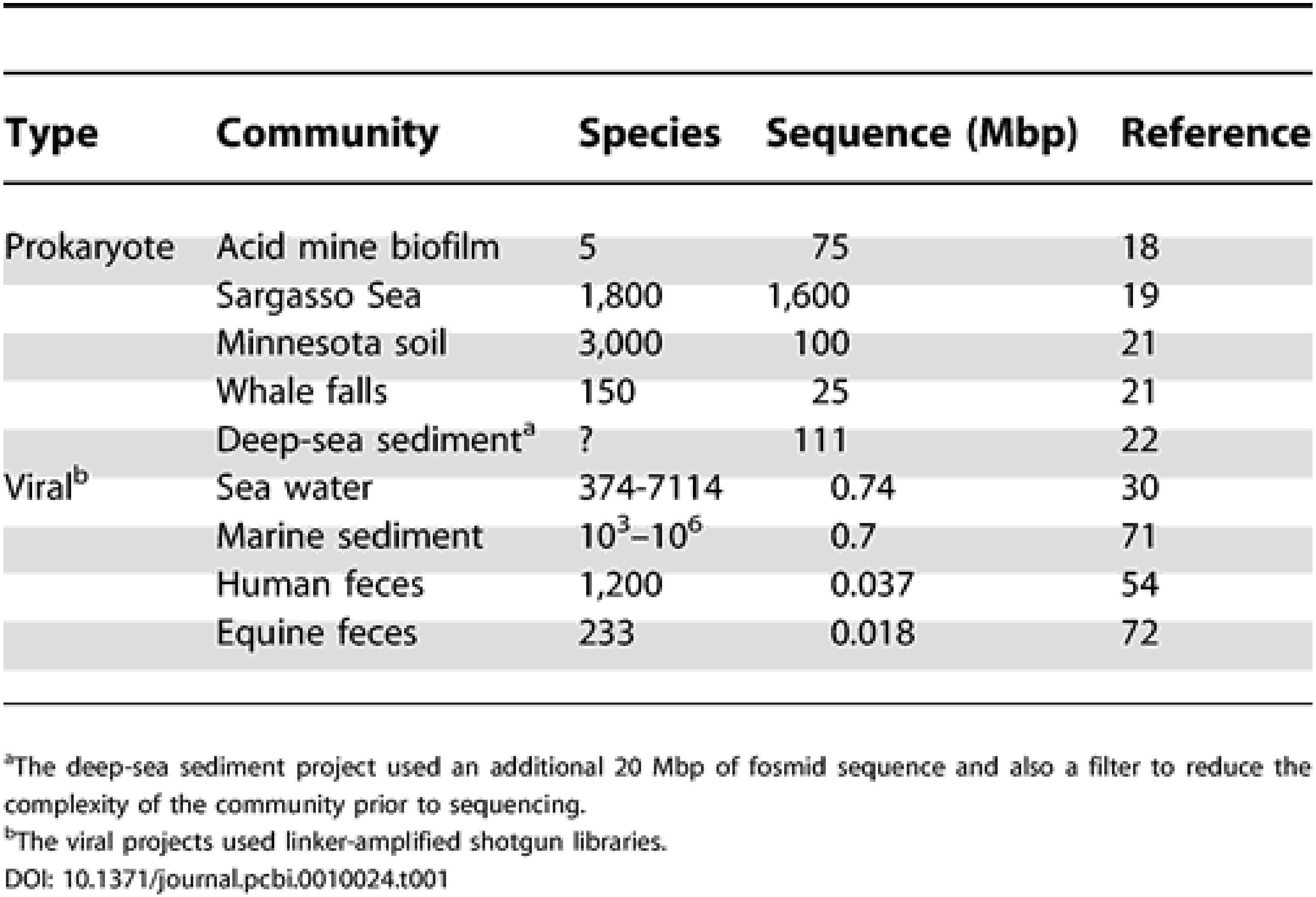
Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- Unlike previous approaches, which typically study a single gene or individual genomes, this approach offers a more global view of the community, allowing us
- to better assess levels of phylogenetic diversity and intraspecies polymorphism,
- study the full gene complement and metabolic pathways in the community,
- and in some cases, reconstruct near-complete genome sequences.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- WGS also has the potential to discover new genes that are too diverged from currently known genes to be amplified with PCR,
- or heterologously expressed in common hosts, and
- is especially important in the case of viral communities because of the lack of a universal gene analogous to *16S*.

Nine shotgun sequencing projects of various communities have been completed to date. The biological insights from these studies have been well-reviewed elsewhere



Type	Community	Species	Sequence (Mbp)	Reference
Prokaryote	Acid mine biofilm	5	75	18
	Sargasso Sea	1,800	1,600	19
	Minnesota soil	3,000	100	21
	Whale falls	150	25	21
	Deep-sea sediment[a]	?	111	22
Viral[b]	Sea water	374-7114	0.74	30
	Marine sediment	10^3–10^6	0.7	71
	Human feces	1,200	0.037	54
	Equine feces	233	0.018	72

[a] The deep-sea sediment project used an additional 20 Mbp of fosmid sequence and also a filter to reduce the complexity of the community prior to sequencing.

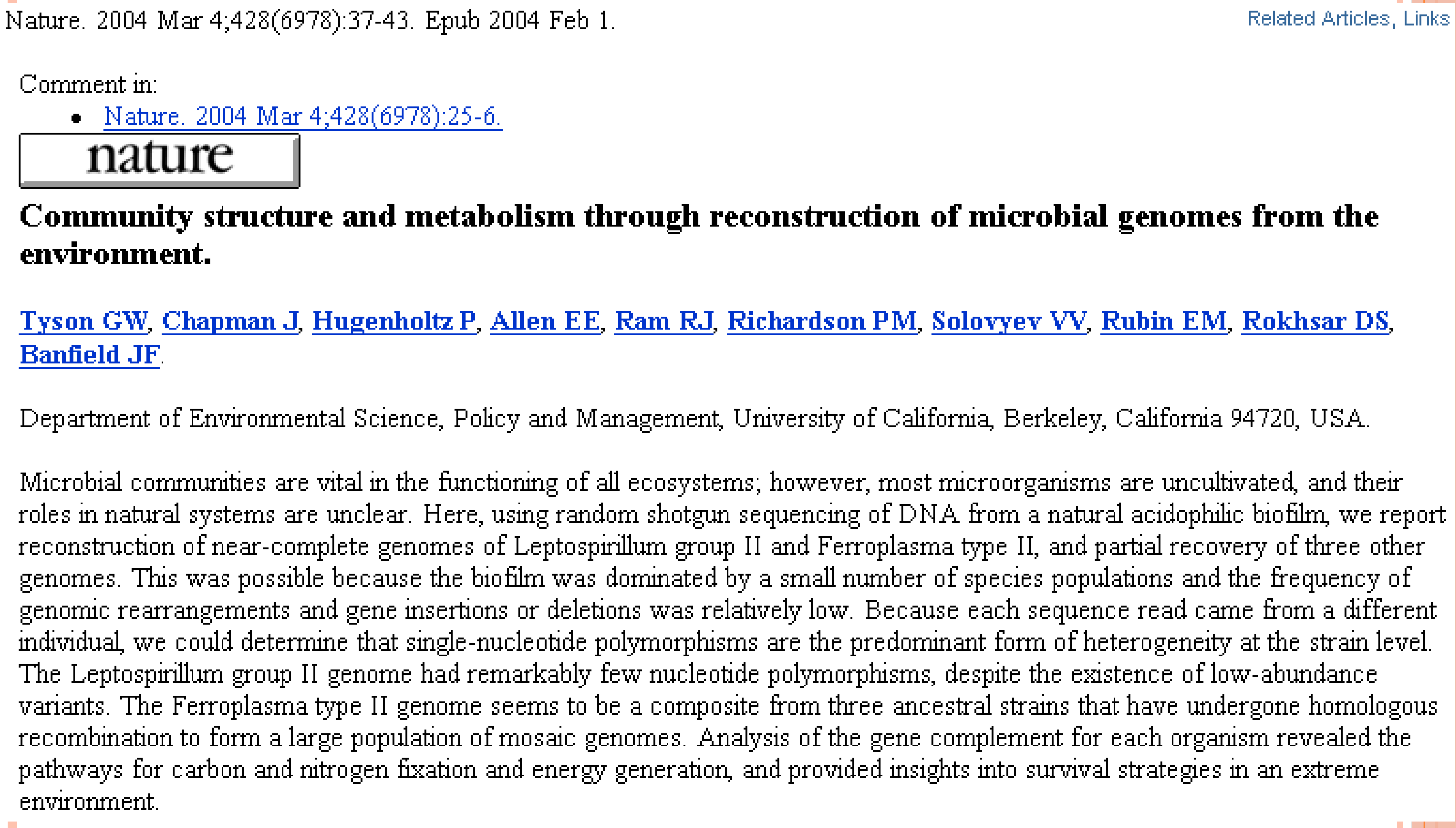
[b] The viral projects used linker-amplified shotgun libraries.

DOI: 10.1371/journal.pcbi.0010024.t001

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- The acid mine biofilm community is an extremely simple model system, consisting of only four dominant species, so a relatively miniscule amount of shotgun sequencing (75 Mbp) was enough to produce two near-complete genome sequences and detailed information about metabolic pathways and strain-level polymorphism.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities



Nature. 2004 Mar 4;428(6978):37-43. Epub 2004 Feb 1.

Comment in:

- Nature. 2004 Mar 4;428(6978):25-6.

Community structure and metabolism through reconstruction of microbial genomes from the environment.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF.

Department of Environmental Science, Policy and Management, University of California, Berkeley, California 94720, USA.

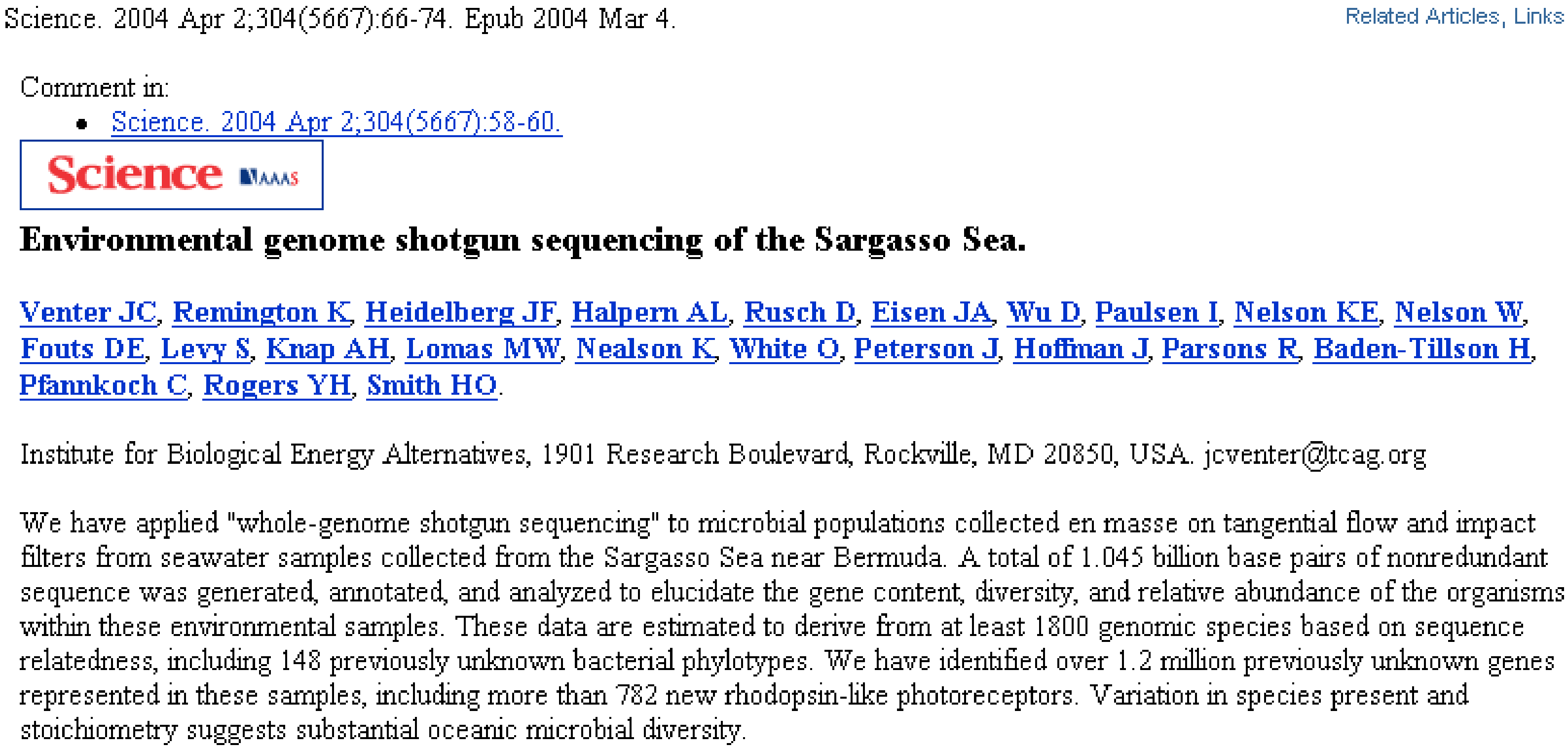
Microbial communities are vital in the functioning of all ecosystems; however, most microorganisms are uncultivated, and their roles in natural systems are unclear. Here, using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of Leptospirillum group II and Ferroplasma type II, and partial recovery of three other genomes. This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low. Because each sequence read came from a different individual, we could determine that single-nucleotide polymorphisms are the predominant form of heterogeneity at the strain level. The Leptospirillum group II genome had remarkably few nucleotide polymorphisms, despite the existence of low-abundance variants. The Ferroplasma type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes. Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- At the other end of the spectrum, the **Sargasso Sea community** is extremely complex, containing more than 1,800 species.
- Nonetheless, with an enormous amount of sequencing (1.6 Gbp), vast amounts of previously unknown diversity were discovered,
- including over 1.2 million new genes,
- 148 new species,
- and numerous new rhodopsin genes.

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

- These results were especially surprising given how well the community had been studied previously, and suggest that equally large amounts of biological diversity await future discovery.



Science. 2004 Apr 2;304(5667):66-74. Epub 2004 Mar 4.

Related Articles, Links

Comment in:

- Science. 2004 Apr 2;304(5667):58-60.

Science AAAS

Environmental genome shotgun sequencing of the Sargasso Sea.

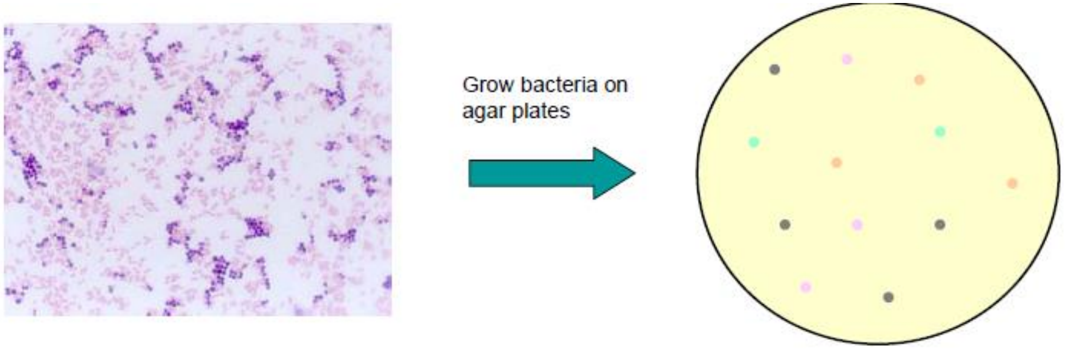
Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO.

Institute for Biological Energy Alternatives, 1901 Research Boulevard, Rockville, MD 20850, USA. jcventer@tcag.org

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

DNA sequencing & microbial profiling

- Traditional microbiology relies on isolation and culture of bacteria
  - Cumbersome and labour intensive process
  - Fails to account for the diversity of microbial life
  - Great plate-count anomaly



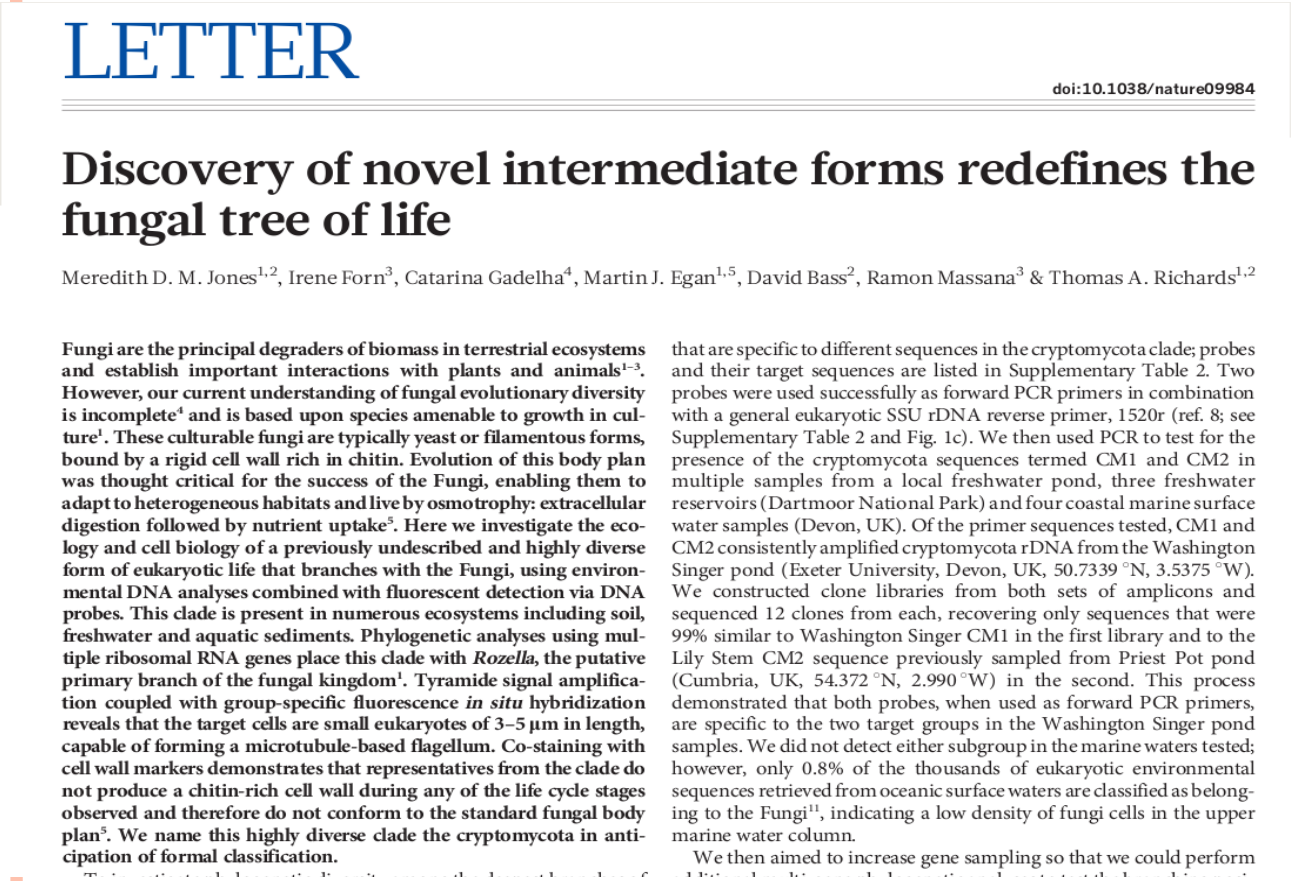
Staley, J. T., and A. Konopka. 1985. Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. **39:**321-346

Why environmental sequencing?

- Only a small proportion of organisms have been grown in culture
- Species do not live in isolation
- Clonal cultures fail to represent the natural environment of a given organism
- Many proteins and protein functions remain undiscovered

Why environmental sequencing?

Estimated 1000 trillion tons of bacterial/archeal life on Earth
Most organisms are difficult to grow in culture



LETTER

doi:10.1038/nature09984

Discovery of novel intermediate forms redefines the fungal tree of life

Meredith D. M. Jones[1,2], Irene Forn[3], Catarina Gadelha[4], Martin J. Egan[1,5], David Bass[2], Ramon Massana[3] & Thomas A. Richards[1,2]

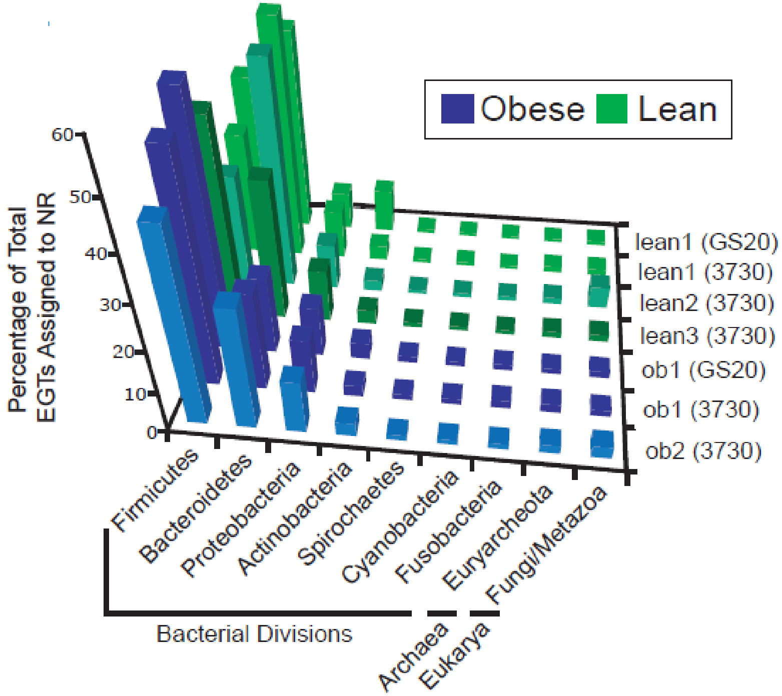
Fungi are the principal degraders of biomass in terrestrial ecosystems and establish important interactions with plants and animals[1–3]. However, our current understanding of fungal evolutionary diversity is incomplete[4] and is based upon species amenable to growth in culture[1]. These culturable fungi are typically yeast or filamentous forms, bound by a rigid cell wall rich in chitin. Evolution of this body plan was thought critical for the success of the Fungi, enabling them to adapt to heterogeneous habitats and live by osmotrophy: extracellular digestion followed by nutrient uptake[5]. Here we investigate the ecology and cell biology of a previously undescribed and highly diverse form of eukaryotic life that branches with the Fungi, using environmental DNA analyses combined with fluorescent detection via DNA probes. This clade is present in numerous ecosystems including soil, freshwater and aquatic sediments. Phylogenetic analyses using multiple ribosomal RNA genes place this clade with *Rozella*, the putative primary branch of the fungal kingdom[1]. Tyramide signal amplification coupled with group-specific fluorescence *in situ* hybridization reveals that the target cells are small eukaryotes of 3–5 μm in length, capable of forming a microtubule-based flagellum. Co-staining with cell wall markers demonstrates that representatives from the clade do not produce a chitin-rich cell wall during any of the life cycle stages observed and therefore do not conform to the standard fungal body plan[5]. We name this highly diverse clade the cryptomycota in anticipation of formal classification.

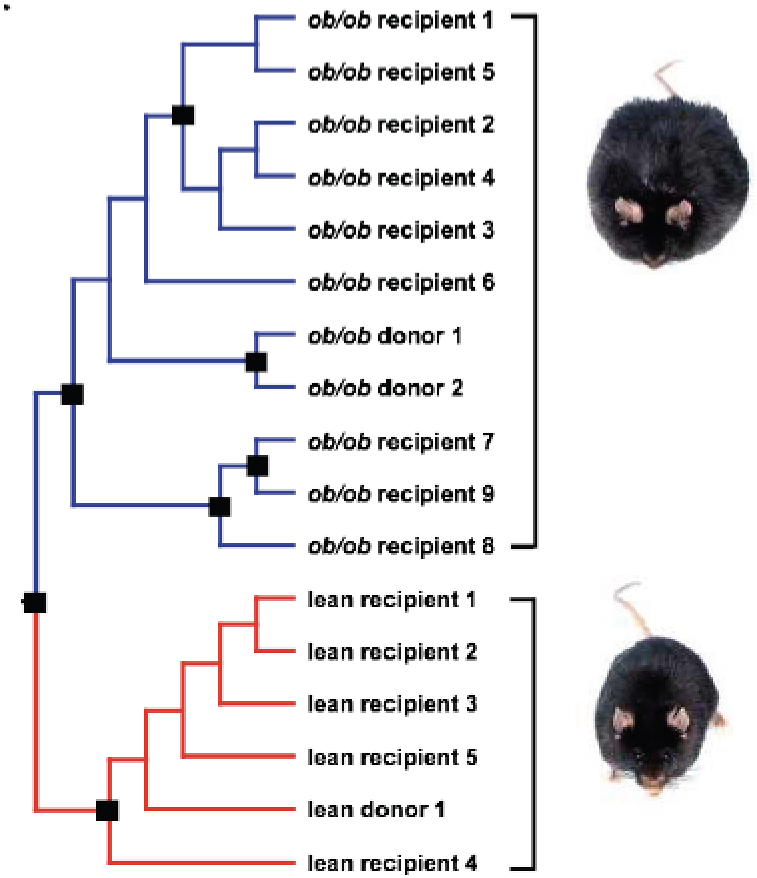
that are specific to different sequences in the cryptomycota clade; probes and their target sequences are listed in Supplementary Table 2. Two probes were used successfully as forward PCR primers in combination with a general eukaryotic SSU rDNA reverse primer, 1520r (ref. 8; see Supplementary Table 2 and Fig. 1c). We then used PCR to test for the presence of the cryptomycota sequences termed CM1 and CM2 in multiple samples from a local freshwater pond, three freshwater reservoirs (Dartmoor National Park) and four coastal marine surface water samples (Devon, UK). Of the primer sequences tested, CM1 and CM2 consistently amplified cryptomycota rDNA from the Washington Singer pond (Exeter University, Devon, UK, 50.7339 °N, 3.5375 °W). We constructed clone libraries from both sets of amplicons and sequenced 12 clones from each, recovering only sequences that were 99% similar to Washington Singer CM1 in the first library and to the Lily Stem CM2 sequence previously sampled from Priest Pot pond (Cumbria, UK, 54.372 °N, 2.990 °W) in the second. This process demonstrated that both probes, when used as forward PCR primers, are specific to the two target groups in the Washington Singer pond samples. We did not detect either subgroup in the marine waters tested; however, only 0.8% of the thousands of eukaryotic environmental sequences retrieved from oceanic surface waters are classified as belonging to the Fungi[11], indicating a low density of fungi cells in the upper marine water column.

We then aimed to increase gene sampling so that we could perform

Jones, M. D. M. et al. Nature (2011).

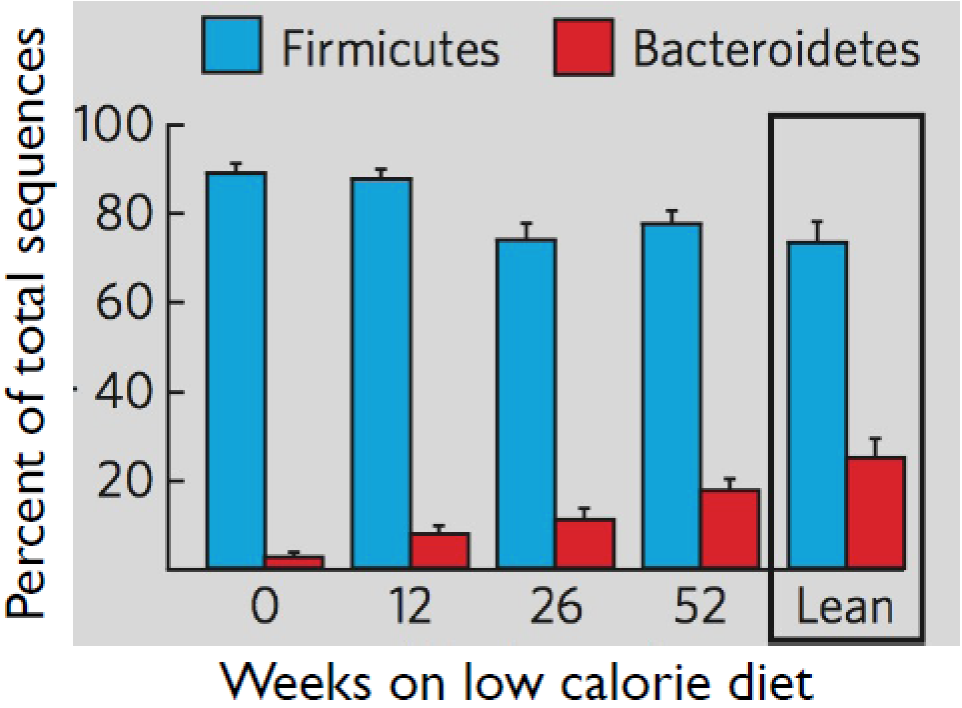
Why environmental sequencing?





Turnbaugh et al. 2006
An obesity associated gut microbiome with increased capacity for energy harvest. Nature 444 1027-1031

Results translate to humans



10x more bacterial cells than human

100-fold more unique genes

Ley et al. 2006
Human Gut Microbiomes associated with obesity.
Nature 444 1022-1023

Overview

- **What is environmental sequencing?**
  - Why?
  - **Methods**
  - Operational Taxonomic Units
  - Measures of diversity
  - Other useful visualisations

DNA sequencing & microbial profiling

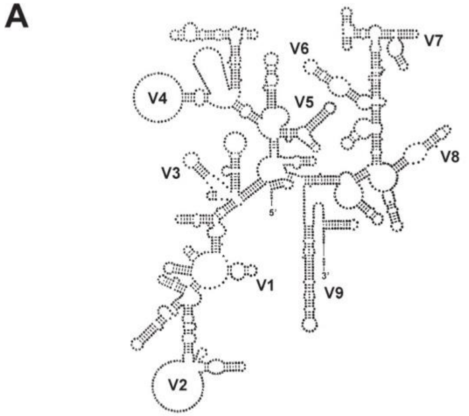
Multiple sequence based options:

- **Sequence tag surveys based on single marker genes**
  - Predominantly 16S rRNA prokaryotes, 18S rRNA for eukaryotes Other genes such as rpoB also be used.
  - Initially done with cloning step and Sanger sequencing (can generate sequences that cover the full-length of the gene)
  - 454 pyrosequencing now the most widely used approach (shorter reads but greater depth)
  - Illumina can also be used with overlapping paired-end reads for even shorter reads but 100x greater depth than 454
  - First trials with PacBio system (1-20kb but only 50,000 seqs/run)

- **Metagenomics**

- **Single-cell genomics**

16S rRNA sequencing



B

V/bp(length)	Species	Strand	Sequence	Tm*
V1/75-53 (22 mer)	*G. lamblia*	α–sense	5' G CGT CCC GGG TGA GCG GGC ATG 3'	35C
	G. muris	α–sense	5' C TTC ATA CCT CGT GTC CGC TTG 3'	26C
	G. ardeae	α–sense	5' T TCG CCG TGC CGC GCC GGC GTG 3'	37C
V3/198-176 (22 mer)	*G. lamblia*	α–sense	5' G CGT CAC GCT GCT CGC CCG CGG 3'	37C
V8/1180-1197 (17 mer)	*G. lamblia*	sense	3' GCC GGC GCC CGC GAG GA 5'	36C
V8/1180-1197 (17 mer)	*G. lamblia*	α–sense	5' CGG CGC CCC GCG AGG AC 3'	36C
	G. ardeae	α–sense	5' GCC TCG CGG ACG TGG TT 3'	29C
	G. muris	α–sense	5' TTT GGT GCA GCT CAA CA 3'	20C

*melting point temperature (Tm) in 2xSSC – 50% formamide

Erlandsen S L et al. J Histochem Cytochem 2005;53:917-927

- 16S rRNA forms part of bacterial ribosomes.

- Contains regions of highly conserved and highly variable sequence.

- Variable sequence can be thought of as a molecular "fingerprint".–can be used to identify bacterial genera and species.

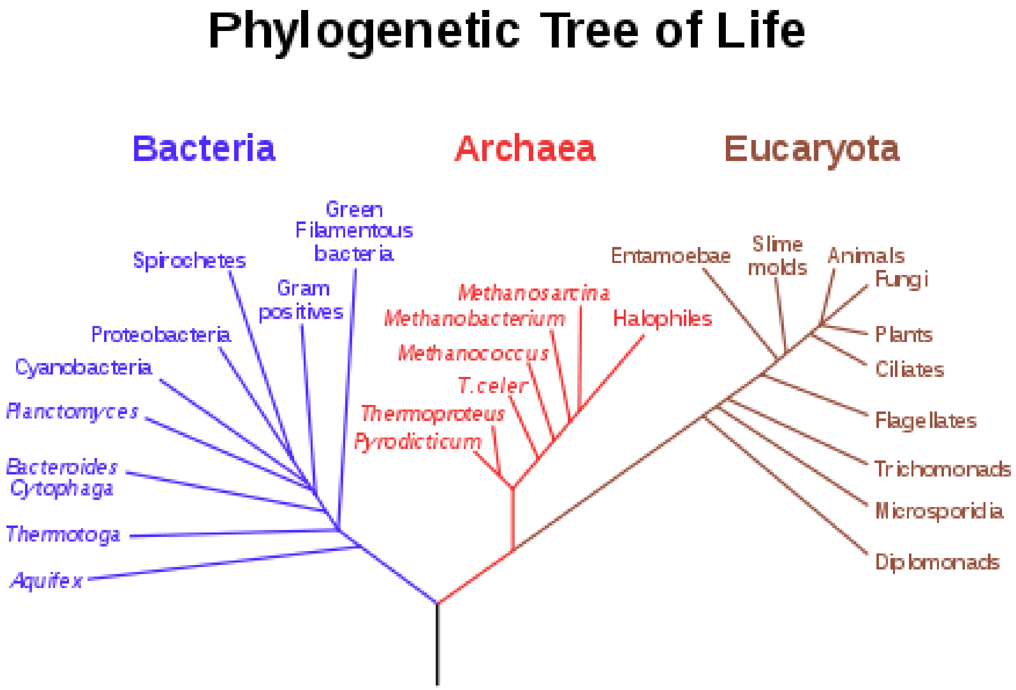
- Large public databases available for comparison.–Ribosomal Database Project currently contains >1.5 million rRNA sequences.

- Conserved regions can be targeted to amplify broad range of bacteria from environmental samples.

- Not quantitative due to copy number variation

Circumvents the need to culture

16S sequencing redefined the tree of life



Woese C, Fox G (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms.". *Proc Natl Acad Sci USA* **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 USA* **87** (12): 4576–9.

Which hyper-variable regions to sequence?

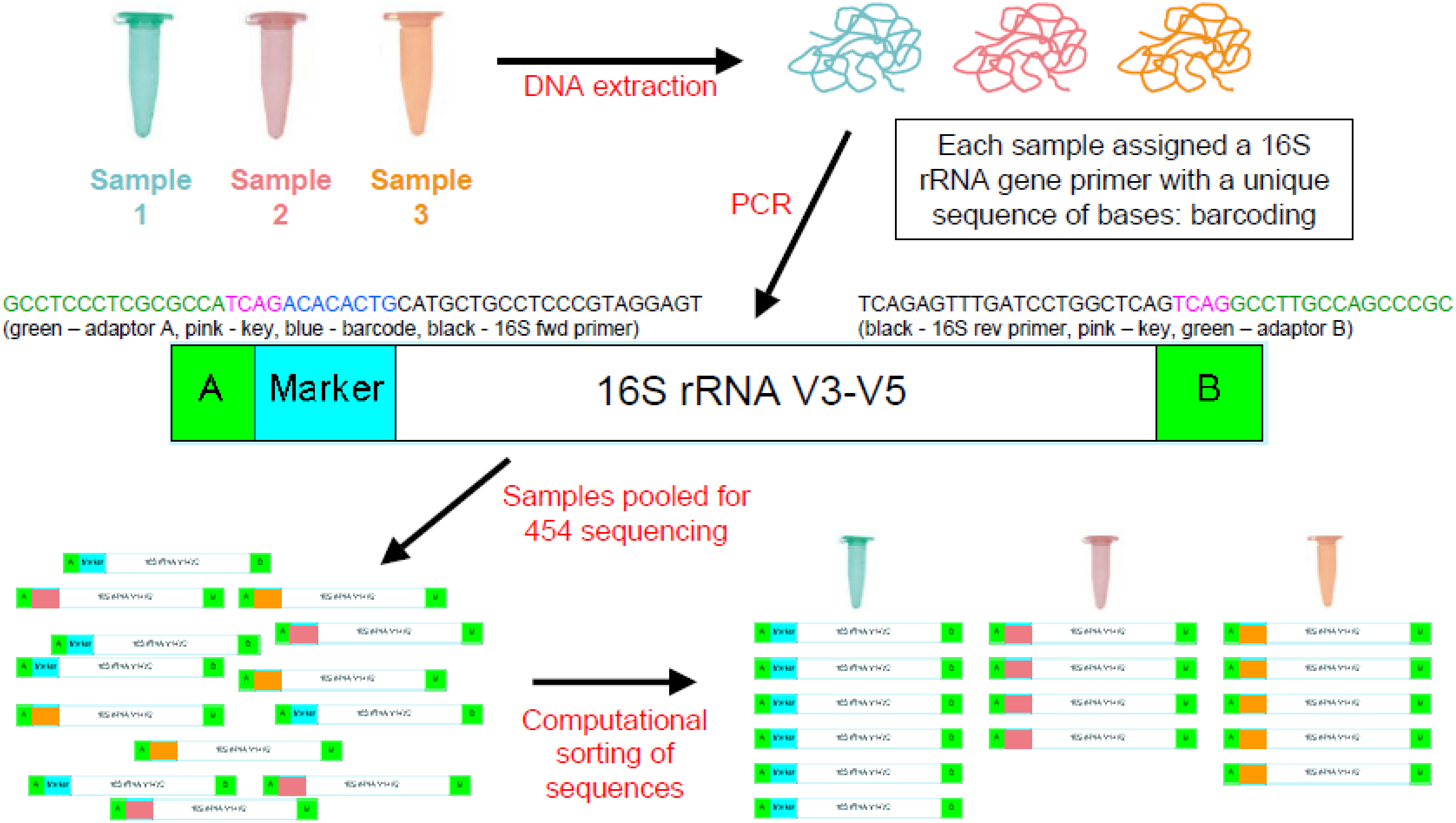
E.coli 16S SSU rRNA hyper-variable regions

Region	Position	# b.p.
V1	69-99	30
V2	137-242	105
V3	338-533	195
V4	576-682	106
V5	822-879	57
V6	967-1046	79
V7	1117-1173	56
V8	1243-1294	51
V9	1435-1465	30

A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria *J Microbiol Methods. 2007 May ; 69(2): 330–339*

A quantitative map of nucleotide substitution rates in bacterial rRNA van der Peer et al *Nucleic Acids Research, 1996, Vol. 24, No. 17 3381–3391*

16S amplicon sequencing



Sample 1

Sample 2

Sample 3

DNA extraction

PCR

Each sample assigned a 16S rRNA gene primer with a unique sequence of bases: barcoding

GCCTCCCTCGCGCCATCAGACACACTGCATGCTGCCTCCCGTAGGAGT
(green – adaptor A, pink - key, blue - barcode, black - 16S fwd primer)

TCAGAGTTTGATCCTGGCTCAGTCAGGCCTTGCCAGCCCGC
(black - 16S rev primer, pink – key, green – adaptor B)

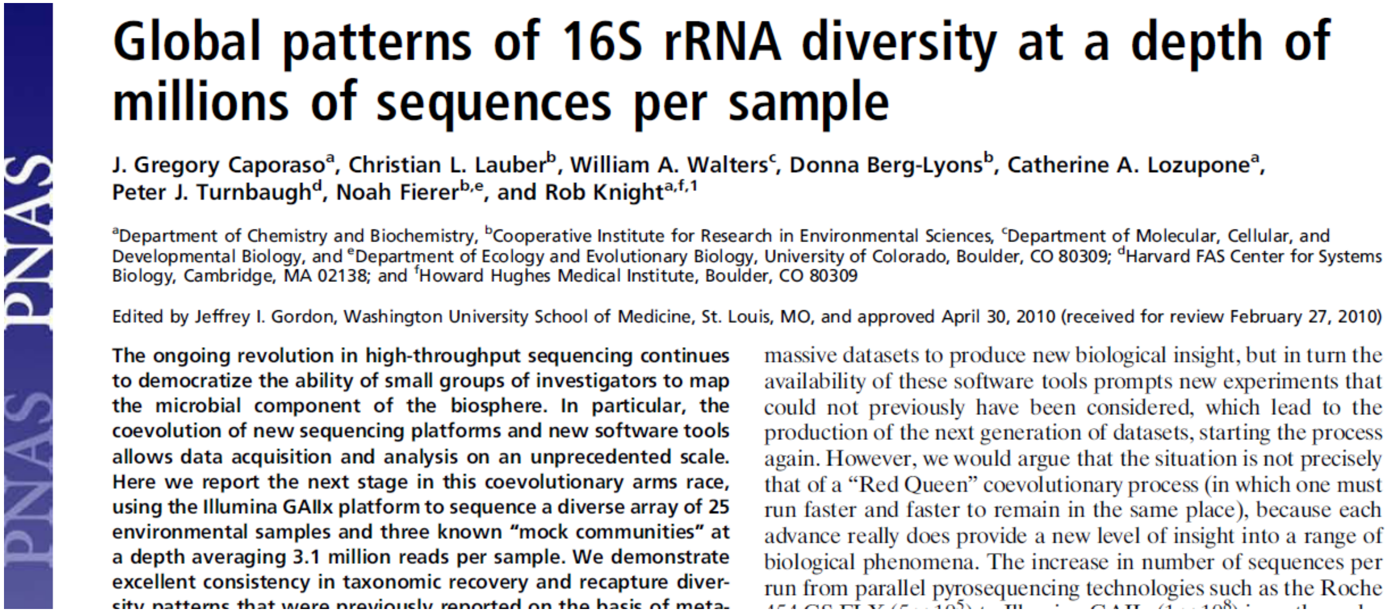
A | Marker | 16S rRNA V3-V5 | B

Samples pooled for 454 sequencing

Computational sorting of sequences

Using overlapping paired-end Illumina reads

- 250bp reads useful for sequencing of individual variable regions (e.g. V3,V6)
- Even single-end reads can be useful
- Enables 3-120 million of reads per sample – 100x more than 454



Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample

J. Gregory Caporaso[a], Christian L. Lauber[b], William A. Walters[c], Donna Berg-Lyons[b], Catherine A. Lozupone[a], Peter J. Turnbaugh[d], Noah Fierer[b,e], and Rob Knight[a,f,1]

[a]Department of Chemistry and Biochemistry, [b]Cooperative Institute for Research in Environmental Sciences, [c]Department of Molecular, Cellular, and Developmental Biology, and [e]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309; [d]Harvard FAS Center for Systems Biology, Cambridge, MA 02138; and [f]Howard Hughes Medical Institute, Boulder, CO 80309

Edited by Jeffrey I. Gordon, Washington University School of Medicine, St. Louis, MO, and approved April 30, 2010 (received for review February 27, 2010)

The ongoing revolution in high-throughput sequencing continues to democratize the ability of small groups of investigators to map the microbial component of the biosphere. In particular, the coevolution of new sequencing platforms and new software tools allows data acquisition and analysis on an unprecedented scale. Here we report the next stage in this coevolutionary arms race, using the Illumina GAIIx platform to sequence a diverse array of 25 environmental samples and three known "mock communities" at a depth averaging 3.1 million reads per sample. We demonstrate excellent consistency in taxonomic recovery and recapture diversity patterns that were previously reported on the basis of meta-

massive datasets to produce new biological insight, but in turn the availability of these software tools prompts new experiments that could not previously have been considered, which lead to the production of the next generation of datasets, starting the process again. However, we would argue that the situation is not precisely that of a "Red Queen" coevolutionary process (in which one must run faster and faster to remain in the same place), because each advance really does provide a new level of insight into a range of biological phenomena. The increase in number of sequences per run from parallel pyrosequencing technologies such as the Roche 454 GS FLX (5 × 10^5) to Illumina GAIIx (1 × 10^8) is

ion torrent
Sequencing for all.™

LONGER READ LENGTHS IMPROVE BACTERIAL IDENTIFICATION USING 16S rRNA GENE SEQUENCING ON THE ION PGM™ SYSTEM

16S rRNA sequencing is a fast, inexpensive profiling technique based on variation in the bacterial 16S ribosomal RNA (rRNA) gene. This method has a wide range of uses, including the characterization of bacteria populations, taxonomical analysis, and species identification. To support diverse projects such as the study of microbes present in foot ulcers and the bioremediation of arsenic-contaminated water, Dr. George Watts (Genomics Shared Service at the University of Arizona Cancer Center, Tucson, AZ) collaborated with Ion Torrent researchers to optimize the amplicon region targeted in the 16S gene (Figure 1) so he could

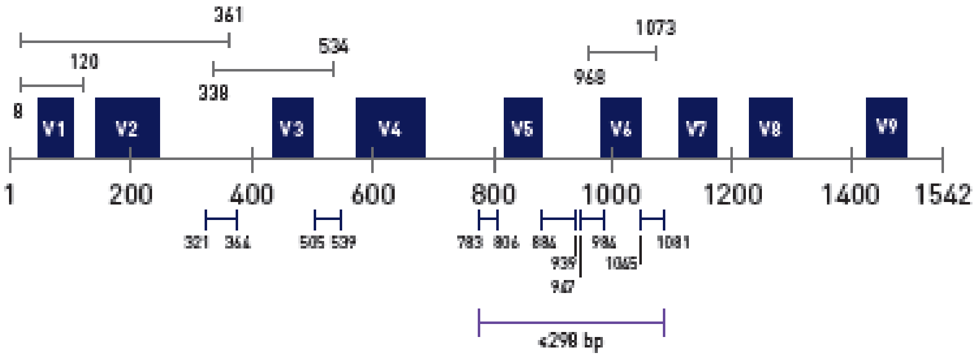


Figure 1. *E. coli* 16S rRNA gene with variable regions illustrated (blue squares). Top panel, amplicons (thin grey bars) targeted in this study. Bottom panel, the thick blue bars indicate conserved regions for primer design due to high coverage for the intervening amplicons for almost all taxa present in the Ribosomal Database Project (RDP) database[26]. The purple bar illustrates an additional hypothetical region that could be used to generate amplicons for 400-base pair-sequencing that would target two variable regions.

Overview

- **What is environmental sequencing?**
  - Why?
  - Methods
  - **Operational Taxonomic Units**
  - Measures of diversity
  - Other useful visualisations

How do we define a species?

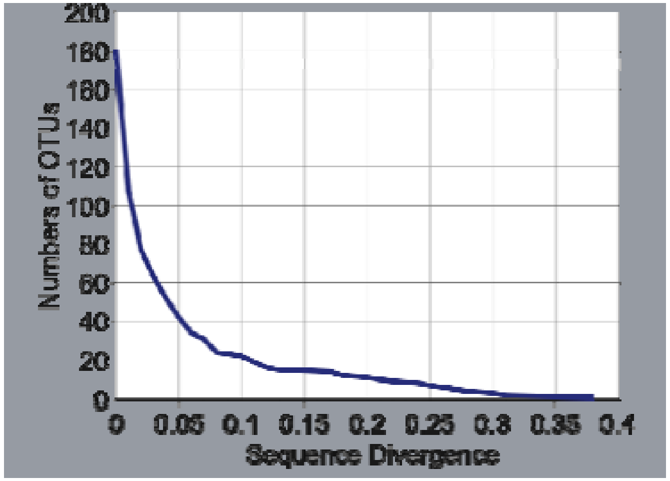
“No single definition has satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species”

Charles Darwin,
On the Origin of Species,
1859

How do we define a species for tag data?

Species concept works for sexually reproducing organisms

- Breaks down when applied to bacteria and fungi
  - Plasmids
  - Horizontal gene transfer
  - Transposons/Viruses
- Operational Taxonomic Unit (OTU)
  - An arbitrary definition of a taxonomic unit based on sequence divergence
  - OTU definitions matter



OTUs definition

OTUs are sequences selected from the reads. The goal is to identify a set of of correct biological sequences.

The concept of an Operational Taxonomic Unit (OTU) was introduced by Peter Sneath and Robert Sokal in the 1960s through a series of books and articles which founded the field of numerical taxonomy (see e.g. Sneath & Sokal: Numerical Taxonomy, W.H. Freeman, 1973).

Their goal was to develop a quantitative strategy for **classifying organisms into groups based on observed characters,** creating a hierarchical classification reflecting the evolutionary relationships between the organisms as faithfully as possible.

Binning tags

Tags may be analysed in one of two ways:

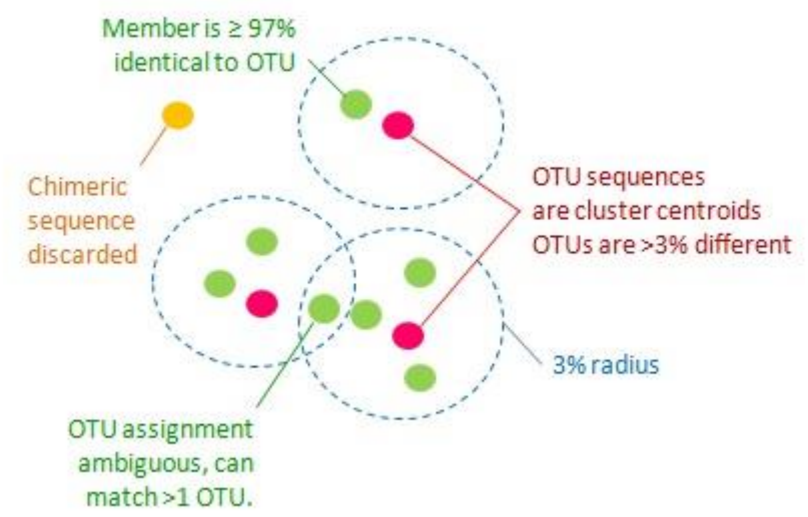
- **Composition-based binning**
  - Relies on comparisons of gross-features to species/genus/families which share these features
    - – GC content
    - – Di/Tri/Tetra/... nucleotide composition (kmer-based frequency comparison)
    - – Codon usage statistics

- **Similarity-based binning**
  - Requires that most sequences in a sample are present in a reference database
    - – Direct comparison of OTU sequence to a reference database
    - – Identity cut-off varies depending on resolution required
      - Genus - 90%
      - Family - 80%
      - Species - 97%
      - Multiple marker genes used for finer sub-strain identification (MLST)
    - – Too stringent cut-off selection will lead to excessive diversity being reported
      - Sequencing errors
      - Sample prep issues

Historical 97% identity threshold
In 16S sequencing, OTUs are typically constructed using an identity threshold of 97%. To the best of my knowledge, the first mention of this threshold is in (Stackebrandt and Goebel 1994).

Stackebrandt and Goebel found that 97% similarity of 16S sequences corresponded approximately to a DNA reassociation value of 70%, which had previously been accepted as a working definition for bacterial species (Wayne et al. 1987).

UPARSE-OTU


Member is ≥ 97% identical to OTU
Chimeric sequence discarded
OTU sequences are cluster centroids OTUs are >3% different
3% radius
OTU assignment ambiguous, can match >1 OTU.

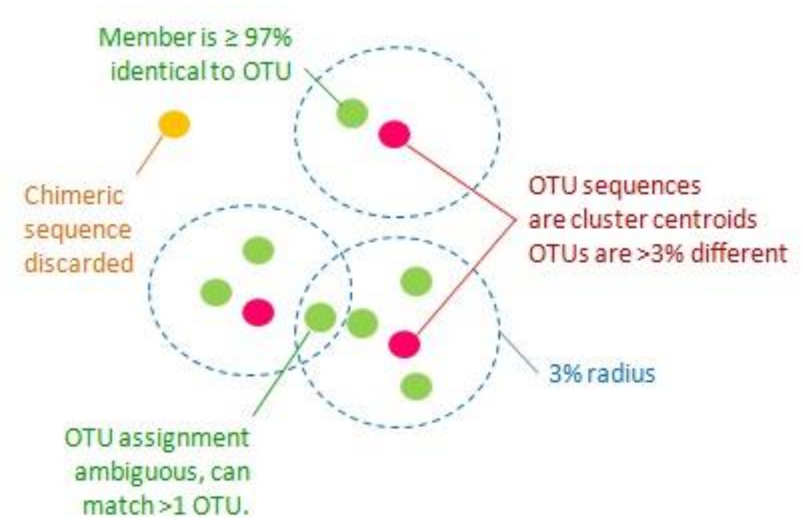
Clustering criteria
The goal of UPARSE-OTU is to identify a set of OTU representative sequences (a subset of the input sequences) satisfying the following criteria.

1. All pairs of OTU sequences should have <97% pair-wise sequence identity.

2. An OTU sequence should be the most abundant within a 97% neighborhood.

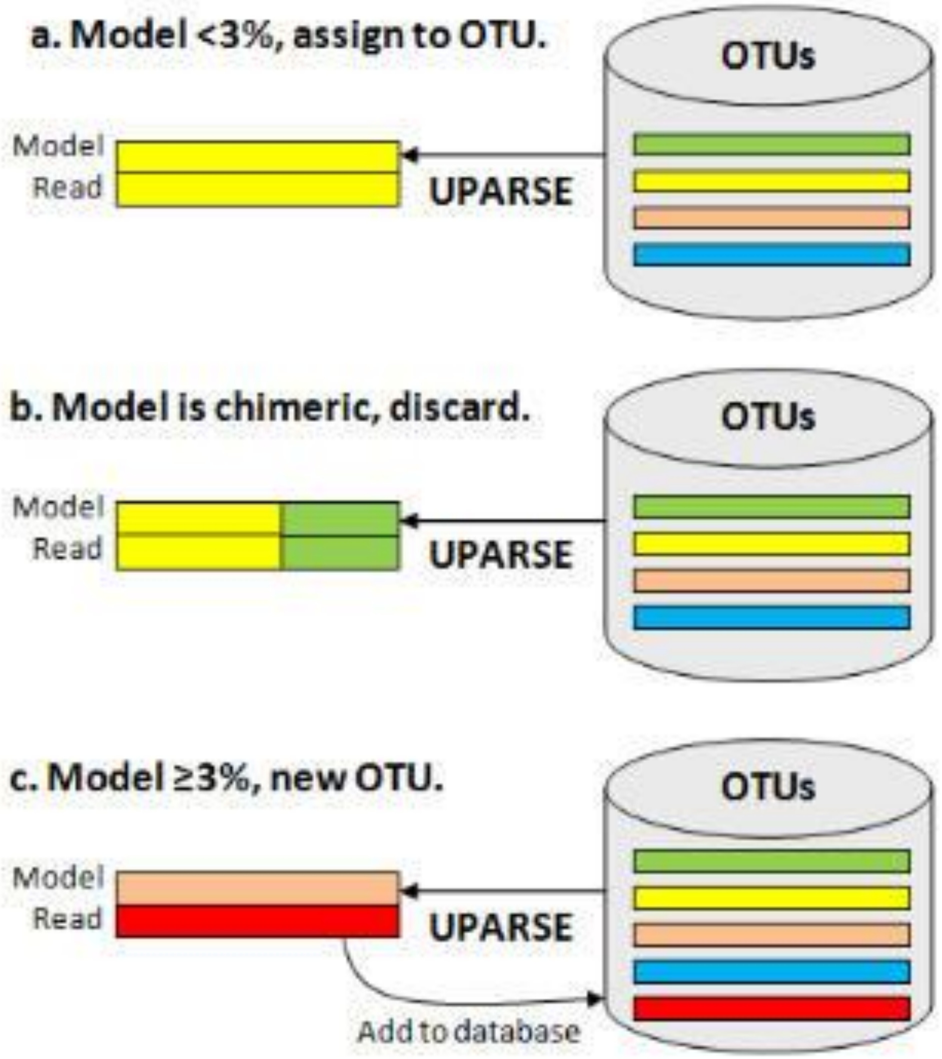
3. Chimeric sequences should be discarded.

4. All non-chimeric input sequences should match at least one OTU with >= 97% identity.


Member is ≥ 97% identical to OTU
Chimeric sequence discarded
OTU sequences are cluster centroids OTUs are >3% different
3% radius
OTU assignment ambiguous, can match >1 OTU.

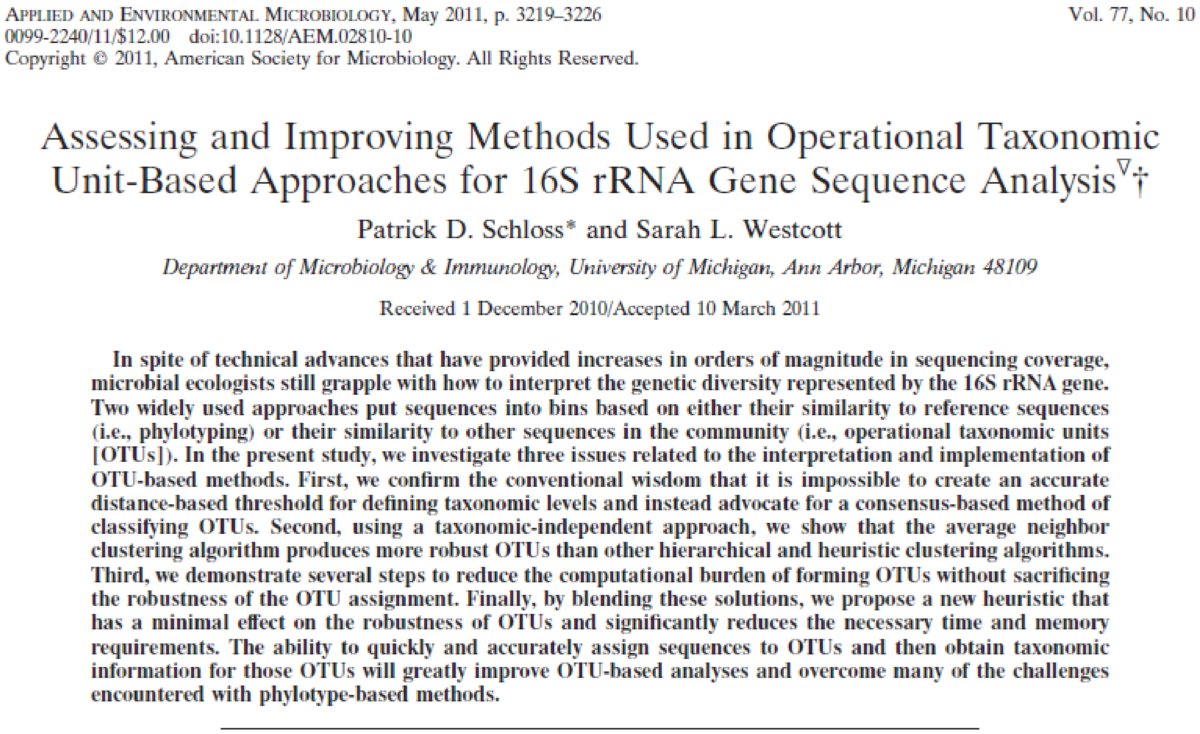
UPARSE-OTU uses a greedy algorithm to find a biologically relevant solution, as follows. Since high-abundance reads are more likely to be correct amplicon sequences, and hence are more likely to be true biological sequences, UPARSE-OTU considers input sequences in order of decreasing abundance.

This means that OTU centroids tend to be selected from the more abundant reads, and hence are more likely to be correct biological sequences.


a. Model <3%, assign to OTU.
Model
Read
UPARSE
OTUs
b. Model is chimeric, discard.
Model
Read
UPARSE
OTUs
c. Model ≥3%, new OTU.
Model
Read
UPARSE
OTUs
Add to database

A word on the importance of clustering algorithms

Average neighbor clustering seems to give the most robust results



Applied and Environmental Microbiology, May 2011, p. 3219–3226 Vol. 77, No. 10
0099-2240/11/$12.00 doi:10.1128/AEM.02810-10
Copyright © 2011, American Society for Microbiology. All Rights Reserved.

Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis ∇ †

Patrick D. Schloss* and Sarah L. Westcott

Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan 48109

Received 1 December 2010/Accepted 10 March 2011

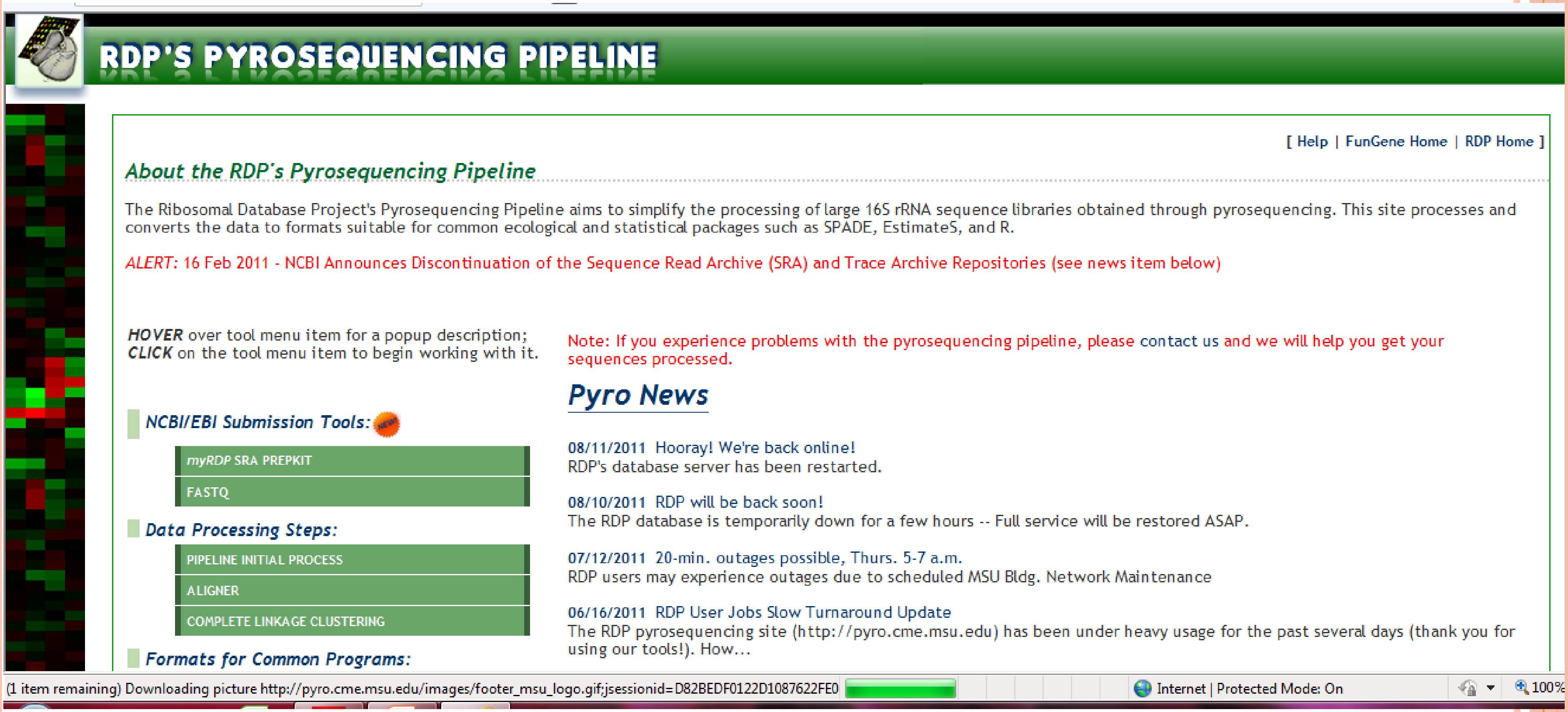
In spite of technical advances that have provided increases in orders of magnitude in sequencing coverage, microbial ecologists still grapple with how to interpret the genetic diversity represented by the 16S rRNA gene. Two widely used approaches put sequences into bins based on either their similarity to reference sequences (i.e., phylotyping) or their similarity to other sequences in the community (i.e., operational taxonomic units [OTUs]). In the present study, we investigate three issues related to the interpretation and implementation of OTU-based methods. First, we confirm the conventional wisdom that it is impossible to create an accurate distance-based threshold for defining taxonomic levels and instead advocate for a consensus-based method of classifying OTUs. Second, using a taxonomic-independent approach, we show that the average neighbor clustering algorithm produces more robust OTUs than other hierarchical and heuristic clustering algorithms. Third, we demonstrate several steps to reduce the computational burden of forming OTUs without sacrificing the robustness of the OTU assignment. Finally, by blending these solutions, we propose a new heuristic that has a minimal effect on the robustness of OTUs and significantly reduces the necessary time and memory requirements. The ability to quickly and accurately assign sequences to OTUs and then obtain taxonomic information for those OTUs will greatly improve OTU-based analyses and overcome many of the challenges encountered with phylotype-based methods.

Software for binning tags

- **Similarity-based binning**
  - Requires that most sequences in a sample are present in a primary or secondary reference database
    - QIIME
    - MEGAN (comparison against Blast NCBI NR)
    - Mothur
    - CARMA (comparison against PFAM)
    - Phymm
    - ARB (linked with Silva database)
    - U-search

Wooley et al. A Primer on Metagenomics, PLoS Computational Biology, Feb 2010, Vol 6(2)

Sequence databases for 16S similarity-based binning



RDP'S PYROSEQUENCING PIPELINE

[Help | FunGene Home | RDP Home]

About the RDP's Pyrosequencing Pipeline

The Ribosomal Database Project's Pyrosequencing Pipeline aims to simplify the processing of large 16S rRNA sequence libraries obtained through pyrosequencing. This site processes and converts the data to formats suitable for common ecological and statistical packages such as SPADE, EstimateS, and R.

ALERT: 16 Feb 2011 - NCBI Announces Discontinuation of the Sequence Read Archive (SRA) and Trace Archive Repositories (see news item below)

HOVER over tool menu item for a popup description;
CLICK on the tool menu item to begin working with it.

NCBI/EBI Submission Tools:

Data Processing Steps:

Formats for Common Programs:

Note: If you experience problems with the pyrosequencing pipeline, please contact us and we will help you get your sequences processed.

Pyro News

08/11/2011 Hooray! We're back online!
RDP's database server has been restarted.

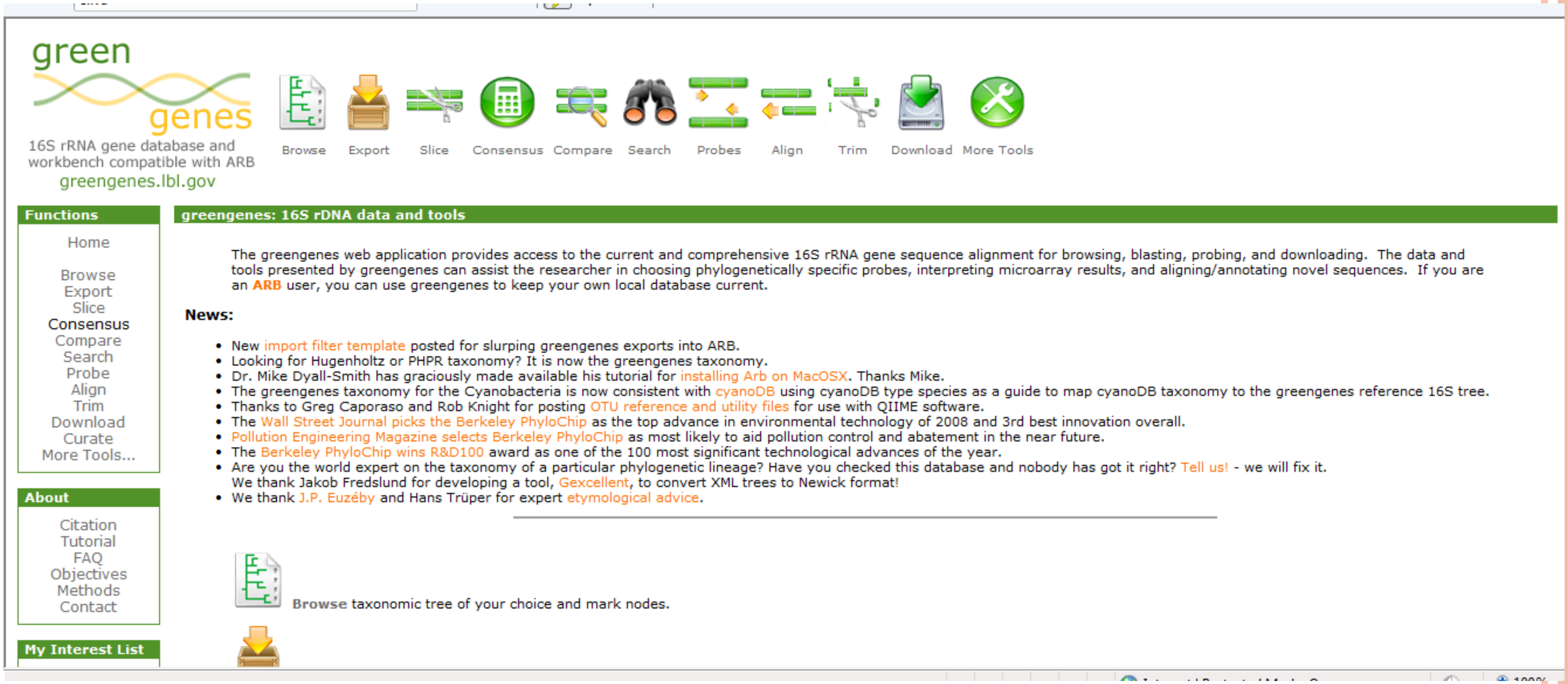
08/10/2011 RDP will be back soon!
The RDP database is temporarily down for a few hours -- Full service will be restored ASAP.

07/12/2011 20-min. outages possible, Thurs. 5-7 a.m.
RDP users may experience outages due to scheduled MSU Bldg. Network Maintenance

06/16/2011 RDP User Jobs Slow Turnaround Update
The RDP pyrosequencing site (http://pyro.cme.msu.edu) has been under heavy usage for the past several days (thank you for using our tools!). How...

(1 item remaining) Downloading picture http://pyro.cme.msu.edu/images/footer_msu_logo.gif;jsessionid=D82BEDF0122D1087622FE0

Sequence databases for 16S similarity-based binning



greengenes: 16S rDNA data and tools

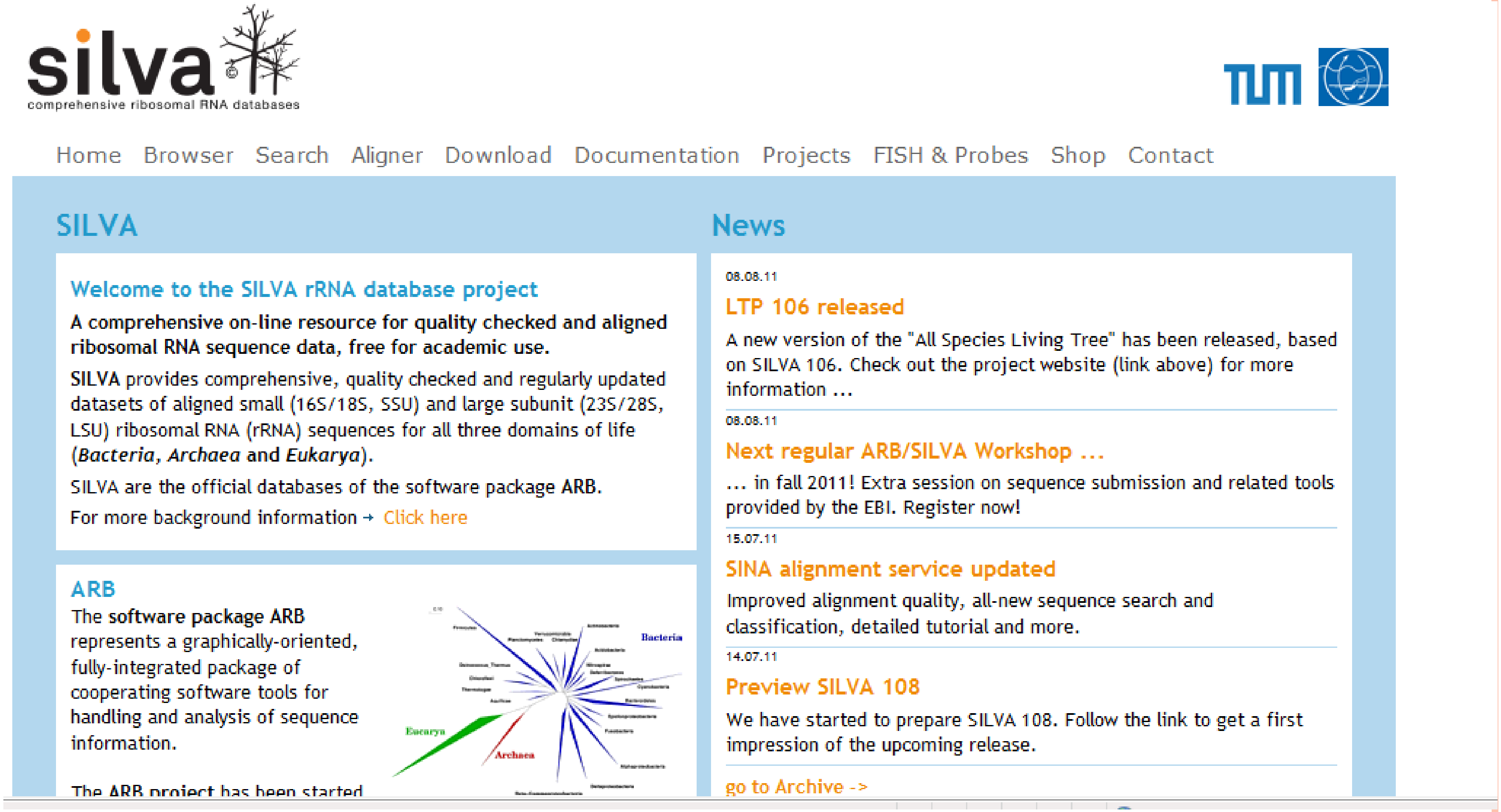
The greengenes web application provides access to the current and comprehensive 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading. The data and tools presented by greengenes can assist the researcher in choosing phylogenetically specific probes, interpreting microarray results, and aligning/annotating novel sequences. If you are an ARB user, you can use greengenes to keep your own local database current.

News:

- New import filter template posted for slurping greengenes exports into ARB.
- Looking for Hugenholtz or PHPR taxonomy? It is now the greengenes taxonomy.
- Dr. Mike Dyall-Smith has graciously made available his tutorial for installing Arb on MacOSX. Thanks Mike.
- The greengenes taxonomy for the Cyanobacteria is now consistent with cyanoDB using cyanoDB type species as a guide to map cyanoDB taxonomy to the greengenes reference 16S tree.
- Thanks to Greg Caporaso and Rob Knight for posting OTU reference and utility files for use with QIIME software.
- The Wall Street Journal picks the Berkeley PhyloChip as the top advance in environmental technology of 2008 and 3rd best innovation overall.
- Pollution Engineering Magazine selects Berkeley PhyloChip as most likely to aid pollution control and abatement in the near future.
- The Berkeley PhyloChip wins R&D100 award as one of the 100 most significant technological advances of the year.
- Are you the world expert on the taxonomy of a particular phylogenetic lineage? Have you checked this database and nobody has got it right? Tell us! - we will fix it.
  We thank Jakob Fredslund for developing a tool, Gexcellent, to convert XML trees to Newick format!
- We thank J.P. Euzéby and Hans Trüper for expert etymological advice.

Browse taxonomic tree of your choice and mark nodes.

Sequence databases for 16S similarity-based binning


silva
comprehensive ribosomal RNA databases
TUM
Home Browser Search Aligner Download Documentation Projects FISH & Probes Shop Contact
Bacteria
Eucarya
Archaea

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data, free for academic use.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

ARB

The **software package ARB** represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.

The ARB project has been started

News

08.08.11

LTP 106 released

A new version of the "All Species Living Tree" has been released, based on SILVA 106. Check out the project website (link above) for more information ...

08.08.11

Next regular ARB/SILVA Workshop ...

... in fall 2011! Extra session on sequence submission and related tools provided by the EBI. Register now!

15.07.11

SINA alignment service updated

Improved alignment quality, all-new sequence search and classification, detailed tutorial and more.

14.07.11

Preview SILVA 108

We have started to prepare SILVA 108. Follow the link to get a first impression of the upcoming release.

go to Archive ->

Overview

- **What is environmental sequencing?**
  - Why?
  - Methods
  - Operational Taxonomic Units
  - **Measures of diversity**
  - Other useful visualisations

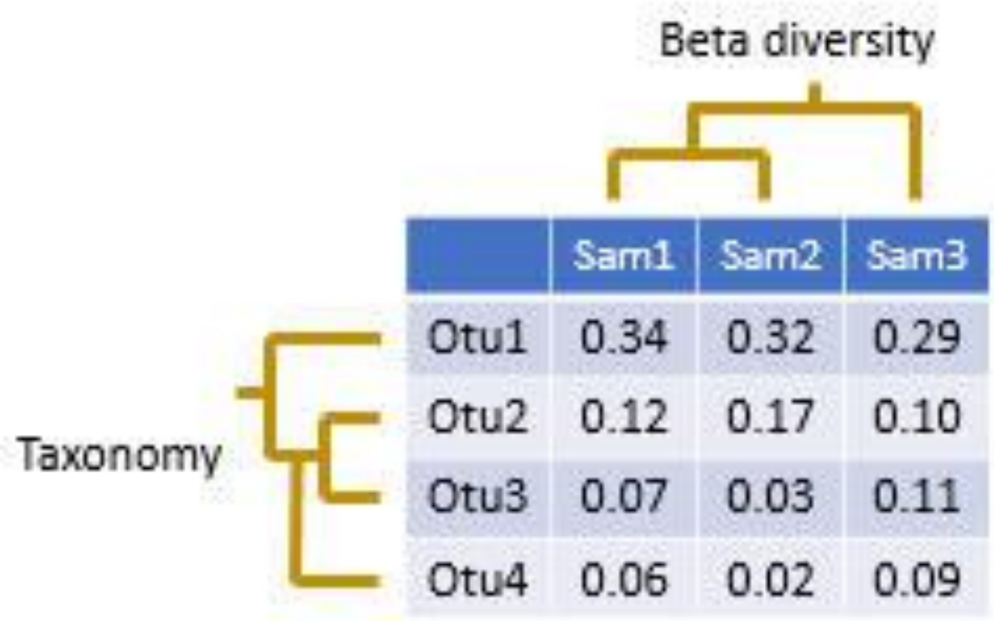
Measuring diversity of OTUs

Two primary measures for sequence based studies:

- Alpha diversity
  - What is there? How much is there?
  - Diversity *within* a sample

- Beta diversity
  - How similar are two samples?
  - Diversity *between* samples

OTU table


Beta diversity
Taxonomy

	Sam1	Sam2	Sam3
Otu1	0.34	0.32	0.29
Otu2	0.12	0.17	0.10
Otu3	0.07	0.03	0.11
Otu4	0.06	0.02	0.09

An OTU table is a matrix that gives the number of reads per sample per OTU. One entry in the table is usually a number of reads, also called a "count", or a frequency in the range 0.0 to 1.0.

It is often assumed that read counts in OTU tables are approximately equivalent to observations of species in traditional ecology. However, interpreting OTU reads counts is actually much more difficult because of biases and errors introduced by PCR and sequencing.

Measuring diversity

Alpha diversity

- Diversity *within* a sample
- Simpson’s diversity index (also Shannon, Chao indexes)
- Gives less weight to rarest species

$$D = 1 - \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)}$$

S is the number of species
N is the total number of organisms
n_i is the number of organisms of species i

Whittaker, R.H. (1972). "Evolution and measurement of species diversity". *Taxon* (International Association for Plant Taxonomy (IAPT)) **21** (2/3): 213–251

Measuring diversity

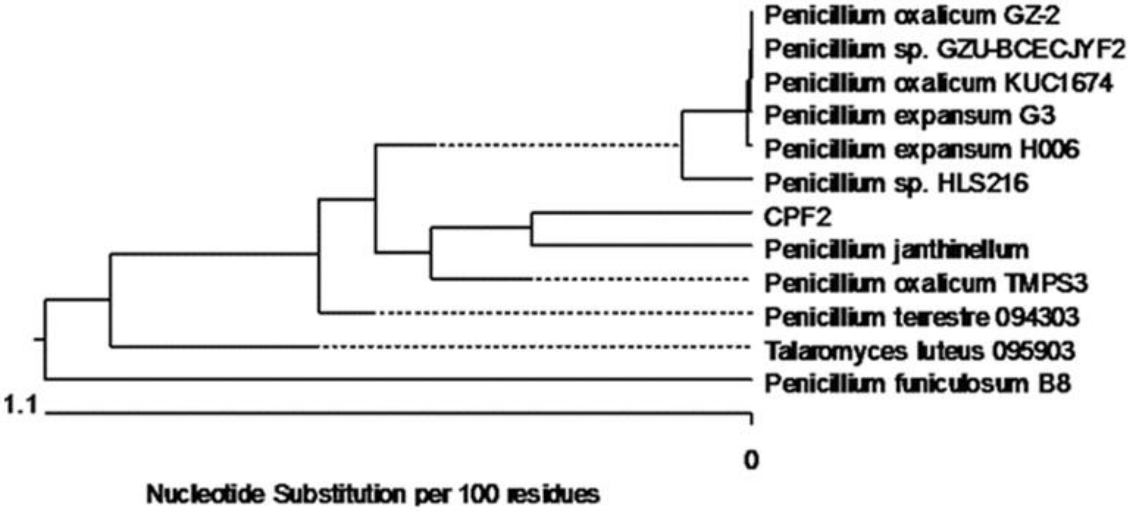
Beta diversity

- Diversity *between* samples
- Sorensen’s index

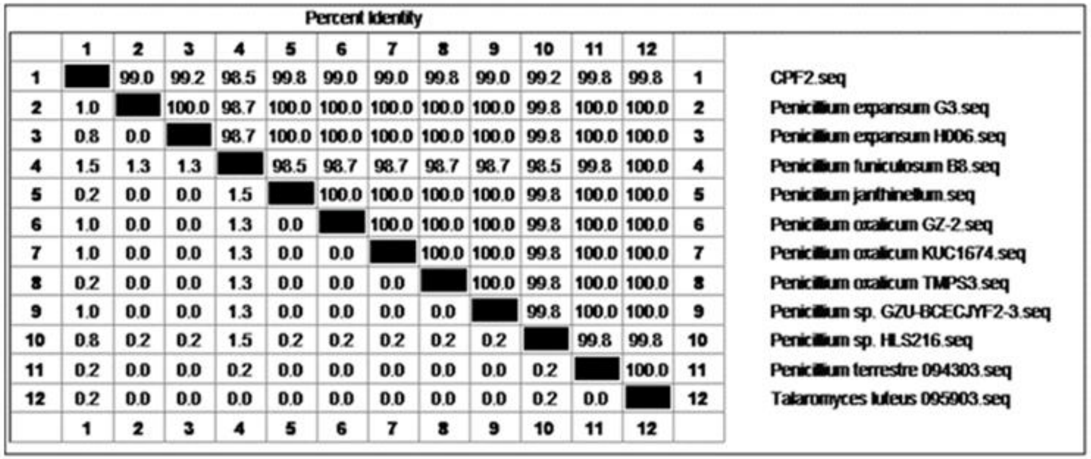
$$\beta = \frac{2c}{S_1 + S_2}$$

S_1 is the number of species in sample 1
S_2 is the number of species in sample 2
c is the number of species present n both samples

Whittaker, R.H. (1972). "Evolution and measurement of species diversity". *Taxon* (International Association for Plant Taxonomy (IAPT)) **21** (2/3): 213–251


Penicillium oxalicum GZ-2
Penicillium sp. GZU-BCECJYF2
Penicillium oxalicum KUC1674
Penicillium expansum G3
Penicillium expansum H006
Penicillium sp. HLS216
CPF2
Penicillium janthinellum
Penicillium oxalicum TMPS3
Penicillium terrestre 094303
Talaromyces luteus 095903
Penicillium funiculosum B8
1.1
0
Nucleotide Substitution per 100 residues

a



Percent Identity

	1	2	3	4	5	6	7	8	9	10	11	12		
1	■	99.0	99.2	98.5	99.8	99.0	99.0	99.8	99.0	99.2	99.8	99.8	1	CPF2.seq
2	1.0	■	100.0	98.7	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	2	Penicillium expansum G3.seq
3	0.8	0.0	■	98.7	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	3	Penicillium expansum H006.seq
4	1.5	1.3	1.3	■	98.5	98.7	98.7	98.7	98.7	98.5	99.8	100.0	4	Penicillium funiculosum B8.seq
5	0.2	0.0	0.0	1.5	■	100.0	100.0	100.0	100.0	99.8	100.0	100.0	5	Penicillium janthinellum.seq
6	1.0	0.0	0.0	1.3	0.0	■	100.0	100.0	100.0	99.8	100.0	100.0	6	Penicillium oxalicum GZ-2.seq
7	1.0	0.0	0.0	1.3	0.0	0.0	■	100.0	100.0	99.8	100.0	100.0	7	Penicillium oxalicum KUC1674.seq
8	0.2	0.0	0.0	1.3	0.0	0.0	0.0	■	100.0	99.8	100.0	100.0	8	Penicillium oxalicum TMPS3.seq
9	1.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	■	99.8	100.0	100.0	9	Penicillium sp. GZU-BCECJYF2-3.seq
10	0.8	0.2	0.2	1.5	0.2	0.2	0.2	0.2	0.2	■	99.8	99.8	10	Penicillium sp. HLS216.seq
11	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	■	100.0	11	Penicillium terrestre 094303.seq
12	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	■	12	Talaromyces luteus 095903.seq
	1	2	3	4	5	6	7	8	9	10	11	12		

b

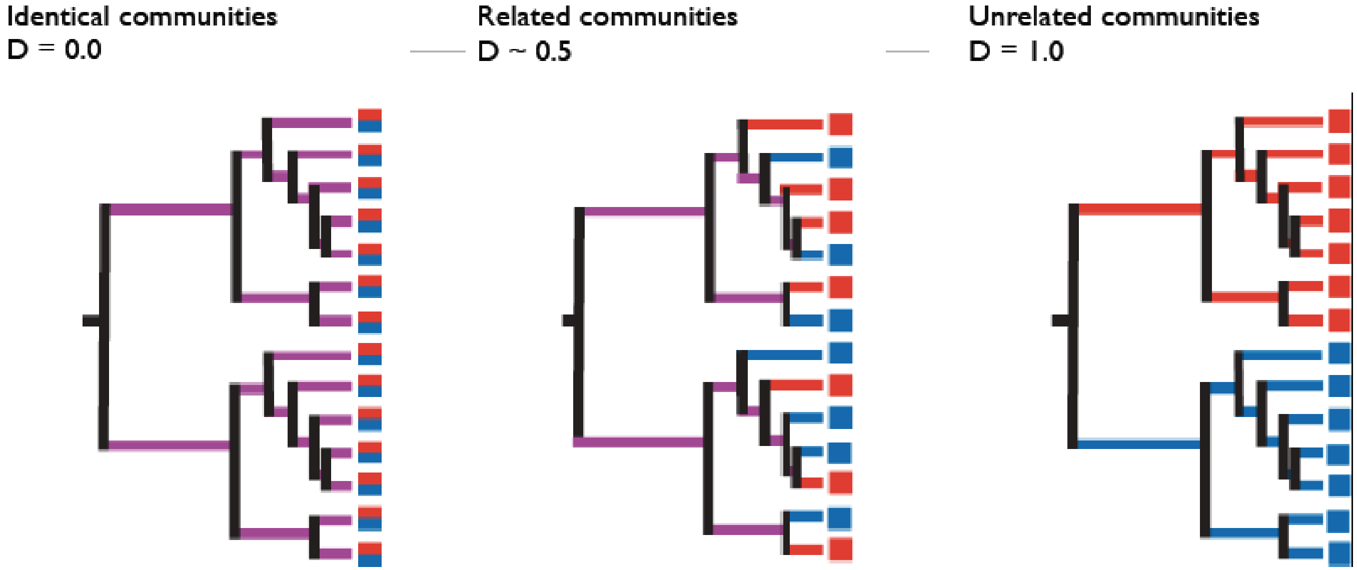
A tree is produced by agglomerative clustering of a distance matrix in tabbed pairs format.

A distance matrix file contains pair-wise distances between a set of sequences, samples, OTUs or other pair-wise comparable objects

Measuring diversity

Beta diversity

- Diversity *between* samples
- Unifrac distance
- Percentage observed branch length unique to either sample



Lozupone and Knight, 2005. Unifrac: A new phylogenetic method for comparing microbial communitieis. Appl Environ Microbiol 71:8228

Overview

- **What is environmental sequencing?**
  - Why?
  - Methods
  - Operational Taxonomic Units
  - Measures of diversity
  - **Other useful visualisations**

Other useful data representations

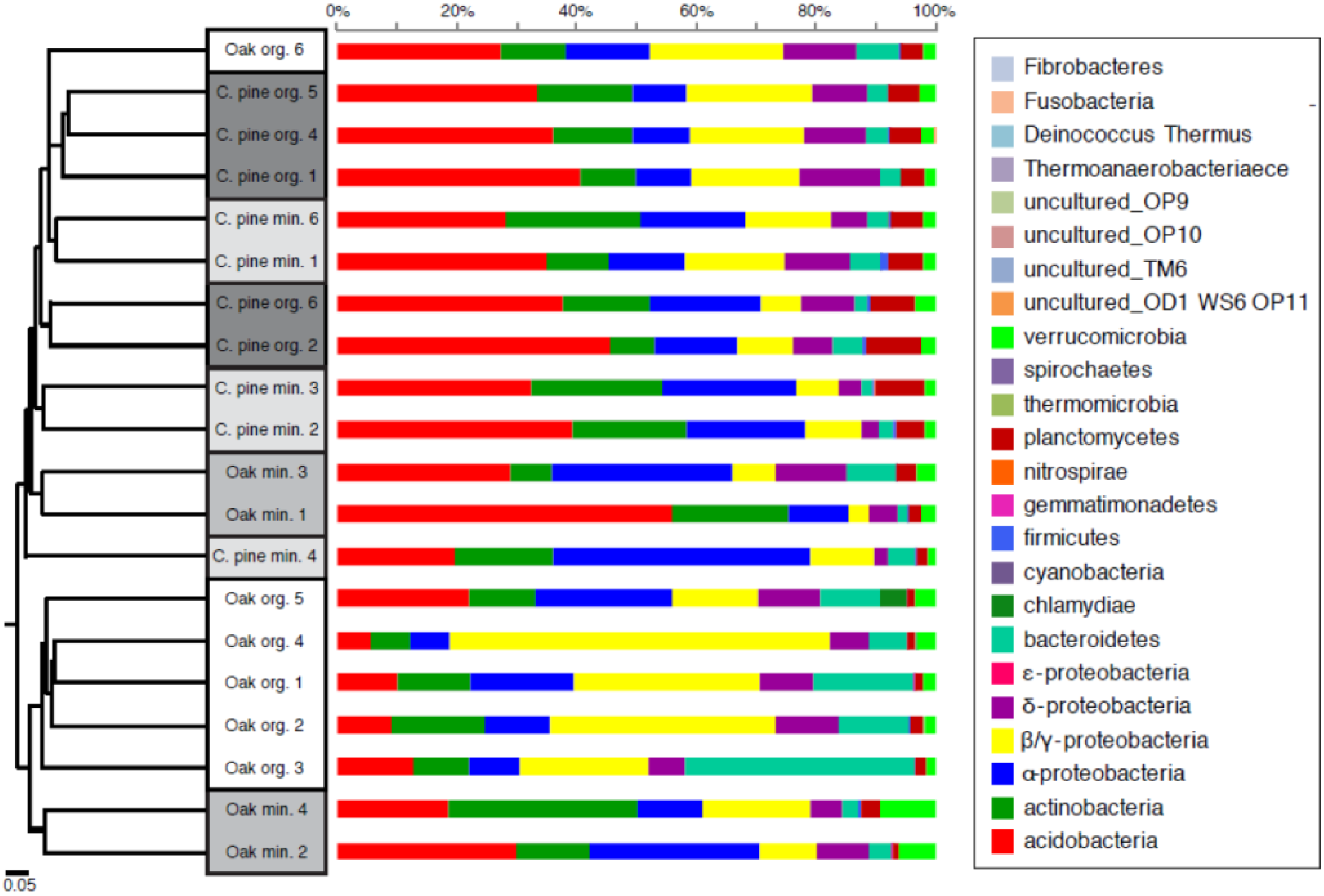
- Simple barcharts
  - What species are present?

- Rarefaction curves
  - How much of a community have we sampled?

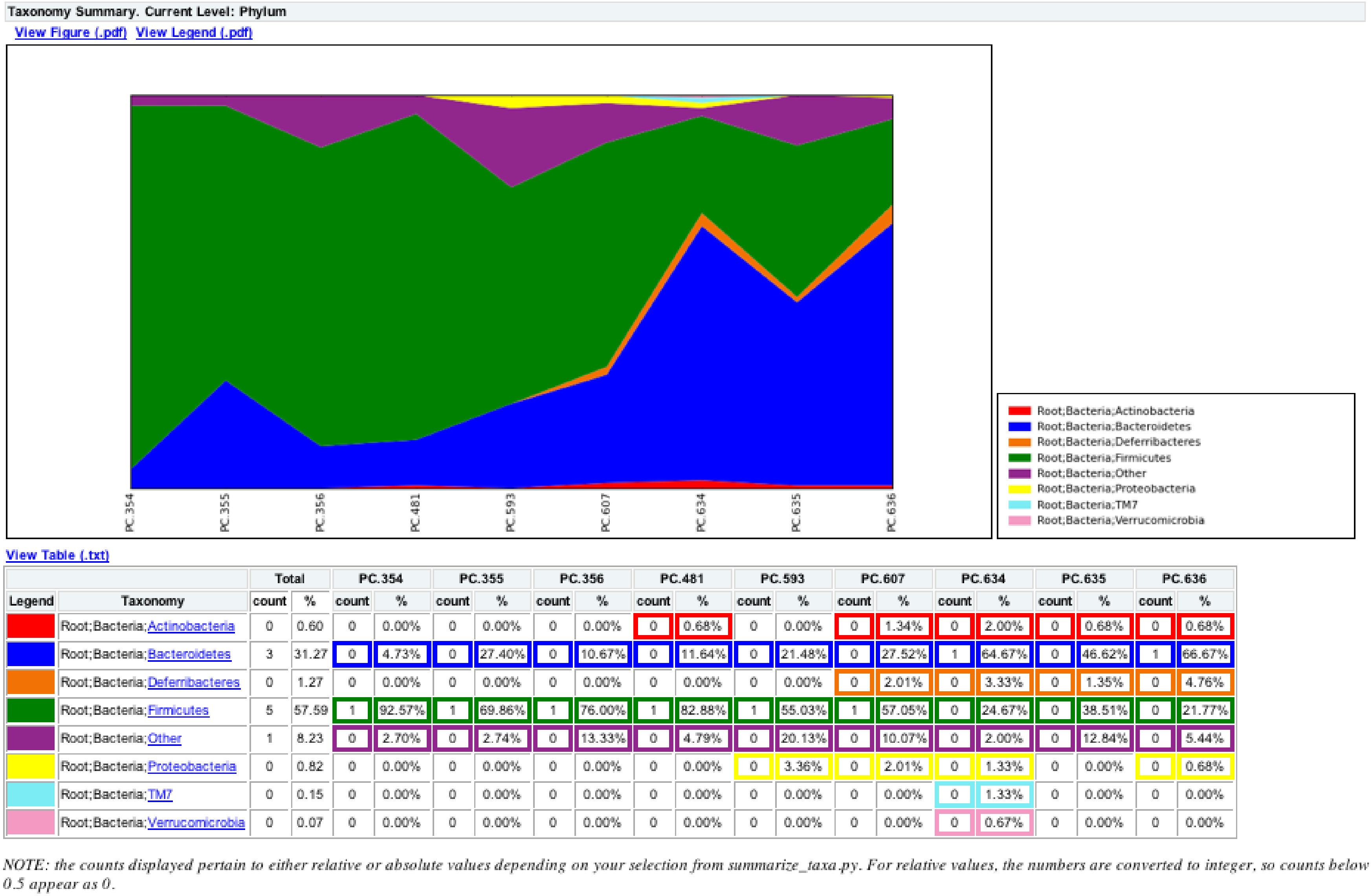
- Principal Component Analysis (PCA)
  - What are the most important factors segregating communities?

- Bootstrapping and jack-knifing
  - How reliable are our measures of diversity?

Simple barcharts



Simple charts



Taxonomy Summary. Current Level: Phylum

View Figure (.pdf) View Legend (.pdf)

PC.354
PC.355
PC.356
PC.481
PC.593
PC.607
PC.634
PC.635
PC.636
Root;Bacteria;Actinobacteria
Root;Bacteria;Bacteroidetes
Root;Bacteria;Deferribacteres
Root;Bacteria;Firmicutes
Root;Bacteria;Other
Root;Bacteria;Proteobacteria
Root;Bacteria;TM7
Root;Bacteria;Verrucomicrobia

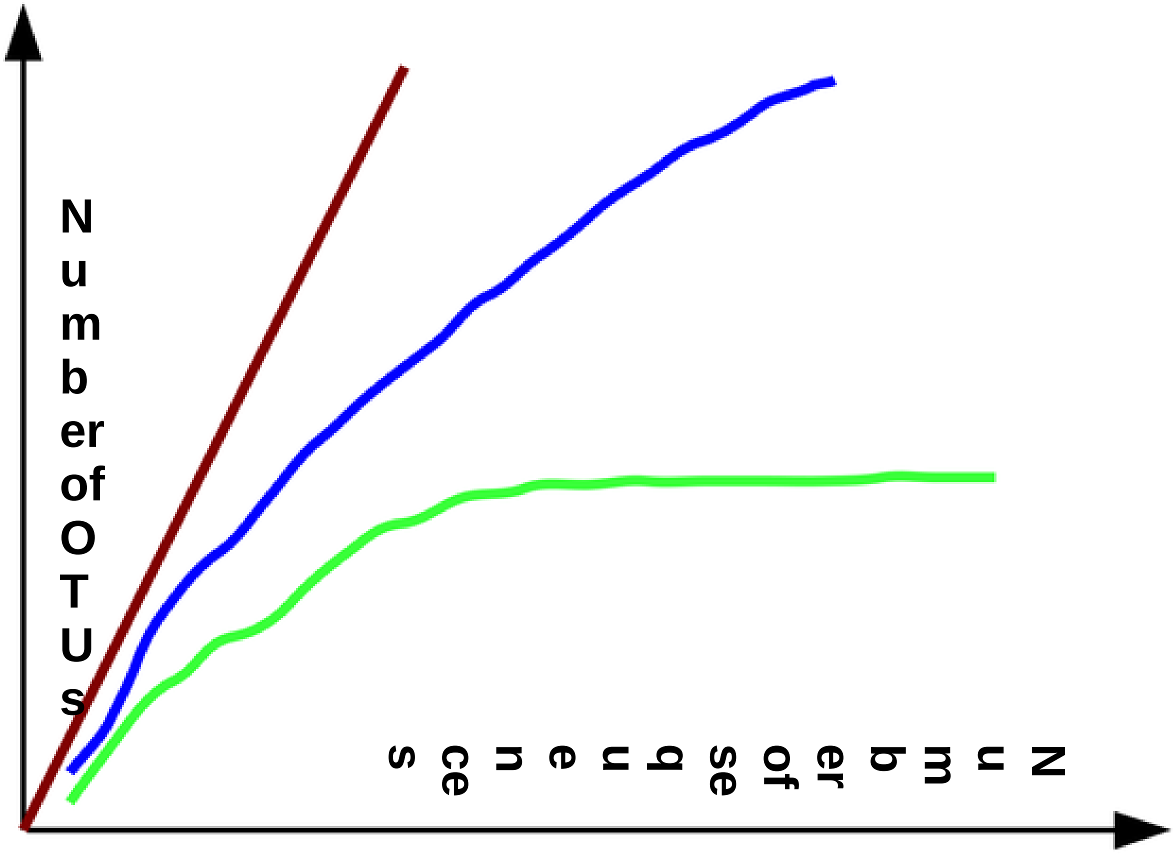
View Table (.txt)

		Total		PC.354		PC.355		PC.356		PC.481		PC.593		PC.607		PC.634		PC.635		PC.636	
Legend	Taxonomy	count	%	count	%	count	%	count	%	count	%	count	%	count	%	count	%	count	%	count	%
🟥	Root;Bacteria;Actinobacteria	0	0.60	0	0.00%	0	0.00%	0	0.00%	0	0.68%	0	0.00%	0	1.34%	0	2.00%	0	0.68%	0	0.68%
🟦	Root;Bacteria;Bacteroidetes	3	31.27	0	4.73%	0	27.40%	0	10.67%	0	11.64%	0	21.48%	0	27.52%	1	64.67%	0	46.62%	1	66.67%
🟧	Root;Bacteria;Deferribacteres	0	1.27	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	2.01%	0	3.33%	0	1.35%	0	4.76%
🟩	Root;Bacteria;Firmicutes	5	57.59	1	92.57%	1	69.86%	1	76.00%	1	82.88%	1	55.03%	1	57.05%	0	24.67%	0	38.51%	0	21.77%
🟪	Root;Bacteria;Other	1	8.23	0	2.70%	0	2.74%	0	13.33%	0	4.79%	0	20.13%	0	10.07%	0	2.00%	0	12.84%	0	5.44%
🟨	Root;Bacteria;Proteobacteria	0	0.82	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	3.36%	0	2.01%	0	1.33%	0	0.00%	0	0.68%
■	Root;Bacteria;TM7	0	0.15	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	1.33%	0	0.00%	0	0.00%
■	Root;Bacteria;Verrucomicrobia	0	0.07	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.67%	0	0.00%	0	0.00%

NOTE: the counts displayed pertain to either relative or absolute values depending on your selection from summarize_taxa.py. For relative values, the numbers are converted to integer, so counts below 0.5 appear as 0.

Rarefaction curves

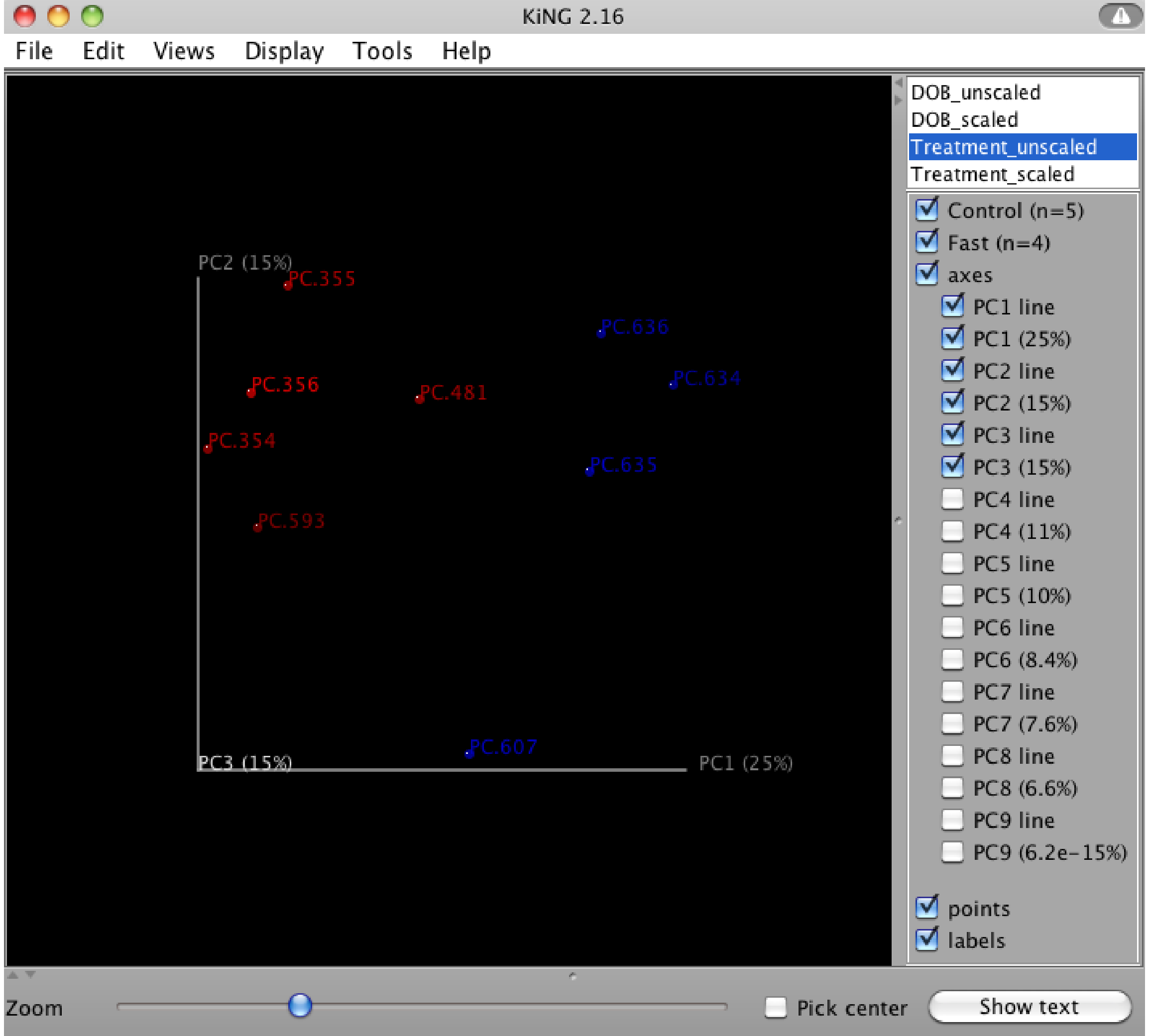
Have we sampled enough of a community to get a true representation?



Adapted from Wooley et al. A Primer on Metagenomics, PLoS Computational Biology, Feb 2010, Vol 6(2)

Principal component analysis

Do samples segregate?



DOB_unscaled
DOB_scaled
Treatment_unscaled
Treatment_scaled

- ☑ Control (n=5)
- ☑ Fast (n=4)
- ☑ axes
  - ☑ PC1 line
  - ☑ PC1 (25%)
  - ☑ PC2 line
  - ☑ PC2 (15%)
  - ☑ PC3 line
  - ☑ PC3 (15%)
  - □ PC4 line
  - □ PC4 (11%)
  - □ PC5 line
  - □ PC5 (10%)
  - □ PC6 line
  - □ PC6 (8.4%)
  - □ PC7 line
  - □ PC7 (7.6%)
  - □ PC8 line
  - □ PC8 (6.6%)
  - □ PC9 line
  - □ PC9 (6.2e−15%)
- ☑ points
- ☑ labels

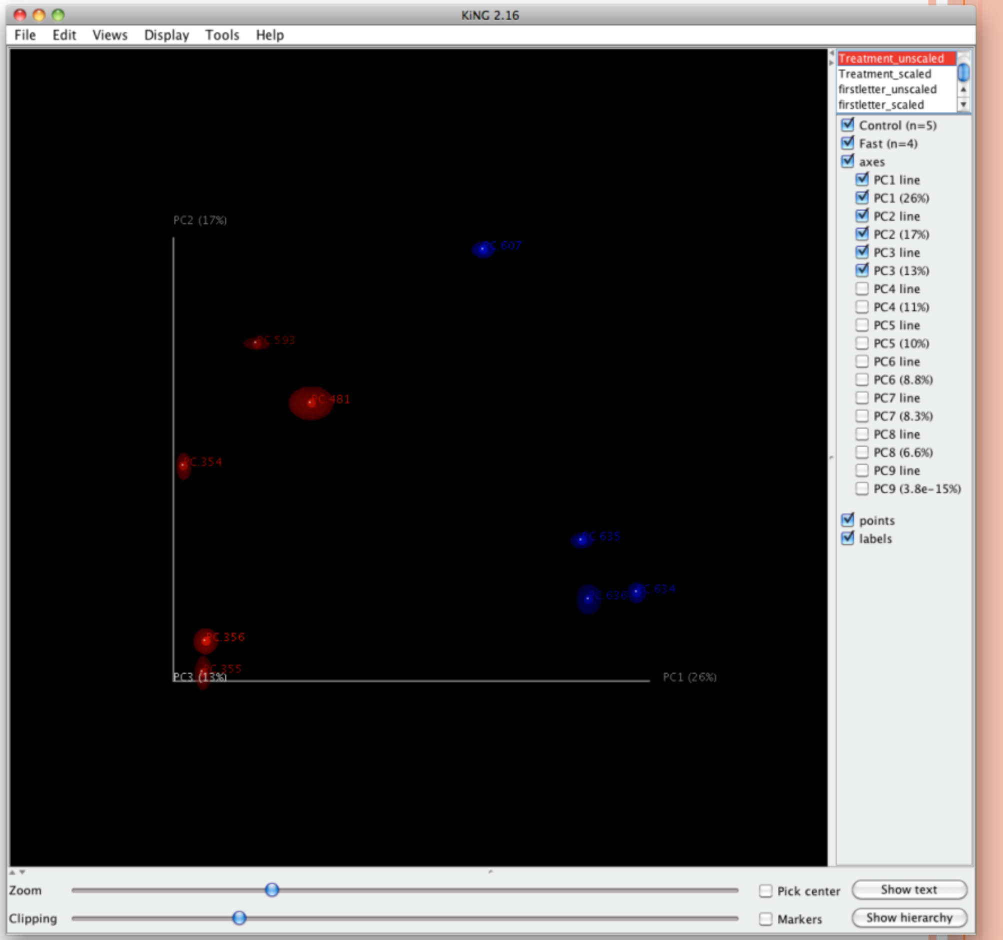
□ Pick center

Jack-knifing

How much uncertainty is there in the clustering and PCA plots?

- Take a subset of your data
- Rerun analysis
- Repeat 100s of times

- Summarize results of 100s of analyses



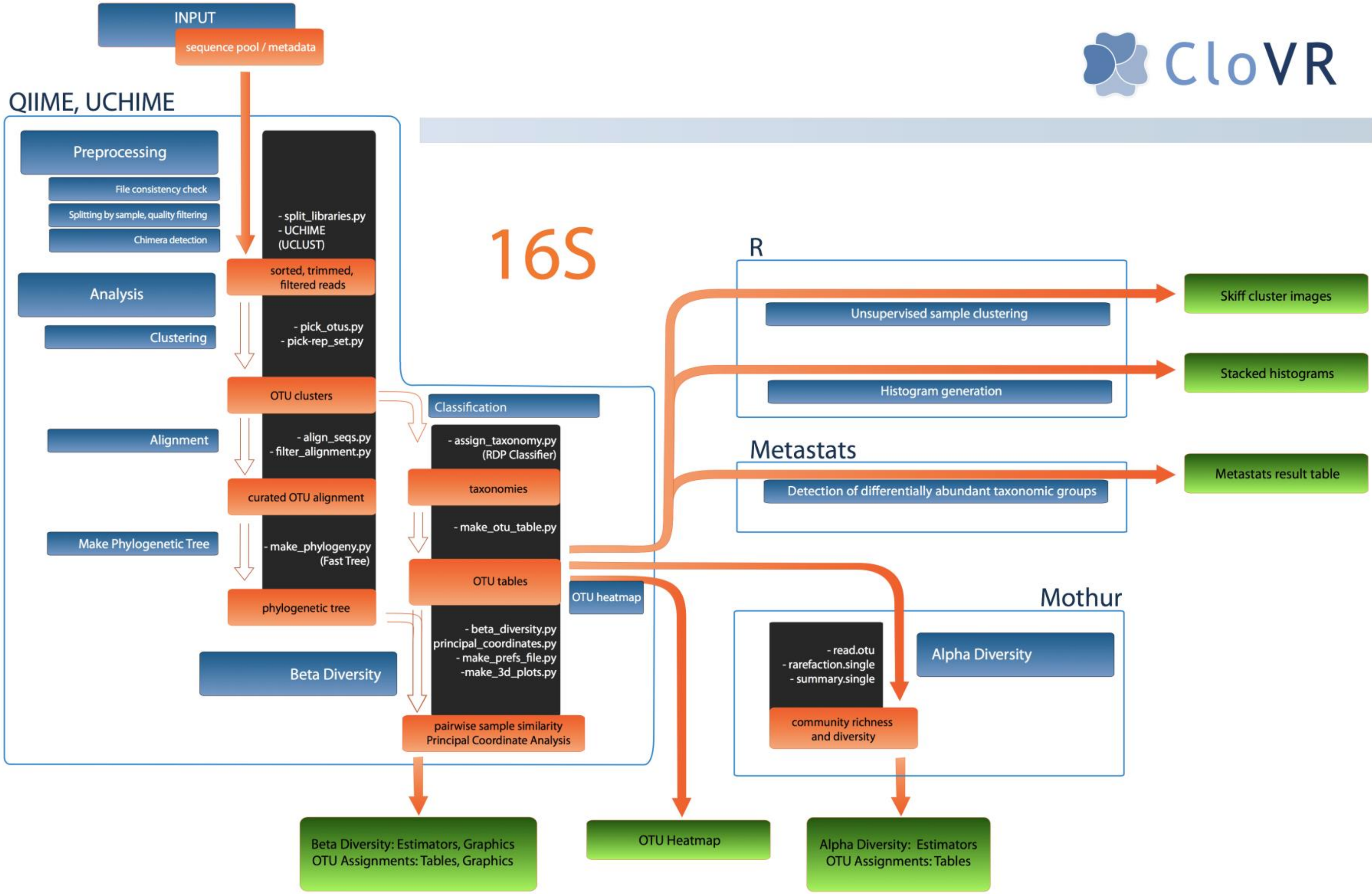
Treatment_unscaled
Treatment_scaled
firstletter_unscaled
firstletter_scaled

☑ Control (n=5)
☑ Fast (n=4)
☑ axes
☑ PC1 line
☑ PC1 (26%)
☑ PC2 line
☑ PC2 (17%)
☑ PC3 line
☑ PC3 (13%)
□ PC4 line
□ PC4 (11%)
□ PC5 line
□ PC5 (10%)
□ PC6 line
□ PC6 (8.8%)
□ PC7 line
□ PC7 (8.3%)
□ PC8 line
□ PC8 (6.6%)
□ PC9 line
□ PC9 (3.8e-15%)

☑ points
☑ labels

□ Pick center
□ Markers

CloVR



16S

INPUT

sequence pool / metadata

QIIME, UCHIME

Preprocessing

File consistency check

Splitting by sample, quality filtering

Chimera detection

- split_libraries.py
- UCHIME (UCLUST)

sorted, trimmed, filtered reads

Analysis

Clustering

- pick_otus.py
- pick-rep_set.py

OTU clusters

Alignment

- align_seqs.py
- filter_alignment.py

curated OTU alignment

Make Phylogenetic Tree

- make_phylogeny.py (Fast Tree)

phylogenetic tree

Beta Diversity

Classification

- assign_taxonomy.py (RDP Classifier)

taxonomies

- make_otu_table.py

OTU tables

OTU heatmap

- beta_diversity.py
principal_coordinates.py
- make_prefs_file.py
-make_3d_plots.py

pairwise sample similarity
Principal Coordinate Analysis

R

Unsupervised sample clustering

Histogram generation

Metastats

Detection of differentially abundant taxonomic groups

Mothur

- read.otu
- rarefaction.single
- summary.single

community richness and diversity

Alpha Diversity

Skiff cluster images

Stacked histograms

Metastats result table

Beta Diversity: Estimators, Graphics
OTU Assignments: Tables, Graphics

OTU Heatmap

Alpha Diversity: Estimators
OTU Assignments: Tables

Overview

- **What is metagenomics?**
  - Why?
  - Case study
  - Assembly, ORFs and Gene finding
  - Annotation

Why metagenomics?

- Tag sequencing can only inform species or strain level classification
- If the species is known and previously sequenced we can have some understanding of the metabolic pathways present due to that organism

- However, most microbes have not been sequenced
- Most have never even been identified

- The depth of sequencing offered by NGS sequencers makes metagenomics feasible
    - Lots of sequences
        - Possible to get a representative sample of all genes present
    - Shorter read length -> hard to assemble

- With current technology the aim is to produce gene catalogues rather than whole genomes
- Limited to prokaryotes

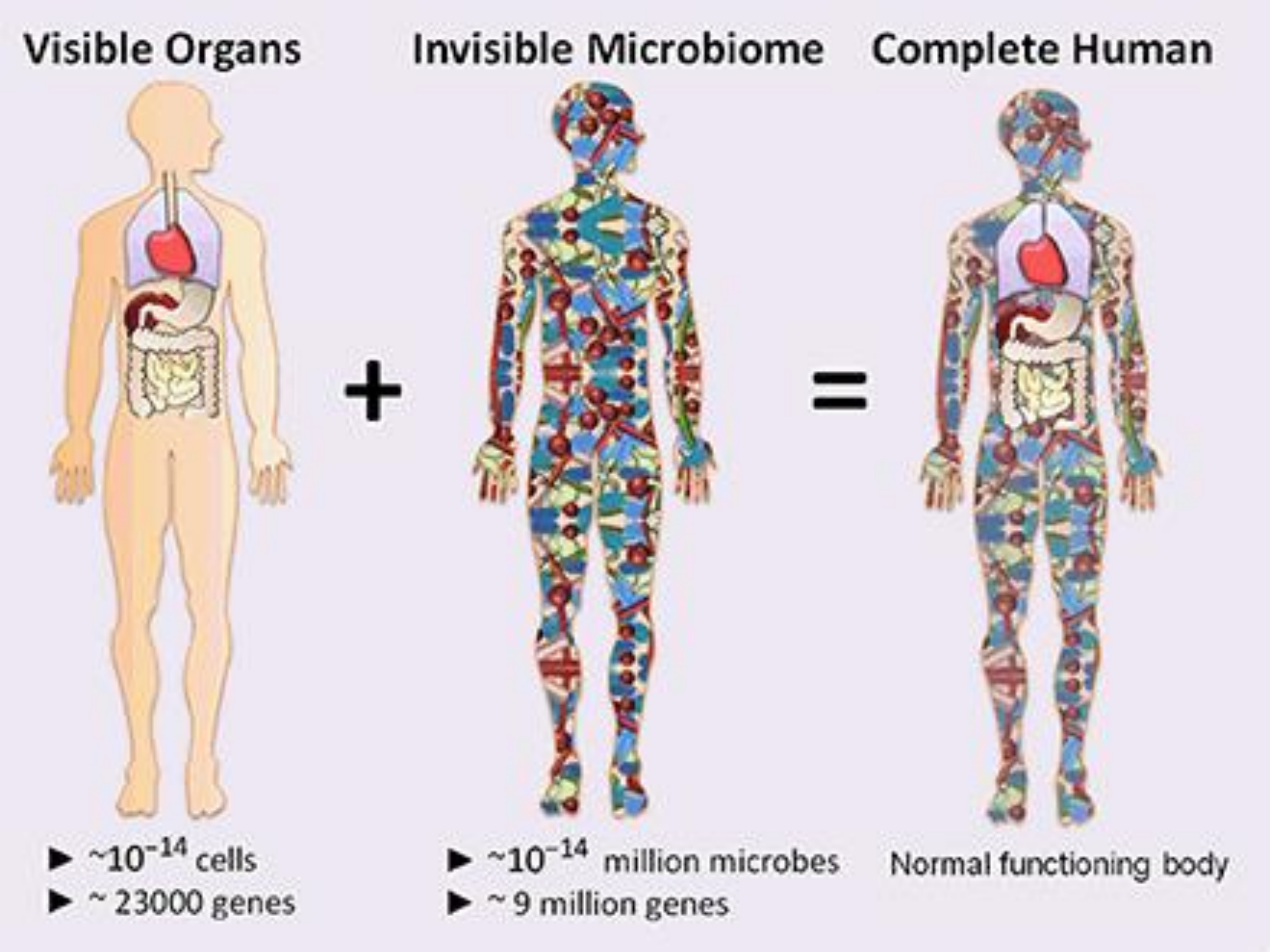
Why metagenomics?

- We contain 100x more bacterial cells than human

- Enivronments of interest
  - Human gut
  - Human skin
  - Human Oral/Nasal and Uritogenetial
  - Chicken gut microbiome
  - Terrabase project (Soil metagenomics)
  - Microbial communities in water (Global Ocean Sampling survey – Venter)
  - Keyboards

- Examine differences between populations (cross-sectional studies)
- Examine changes over time in a single population (longitudinal study)

- Human Microbiome Project
- MetaHIT project


Visible Organs
Invisible Microbiome
Complete Human
+
=
► ~10^{-14} cells
► ~ 23000 genes
► ~10^{-14} million microbes
► ~ 9 million genes
Normal functioning body

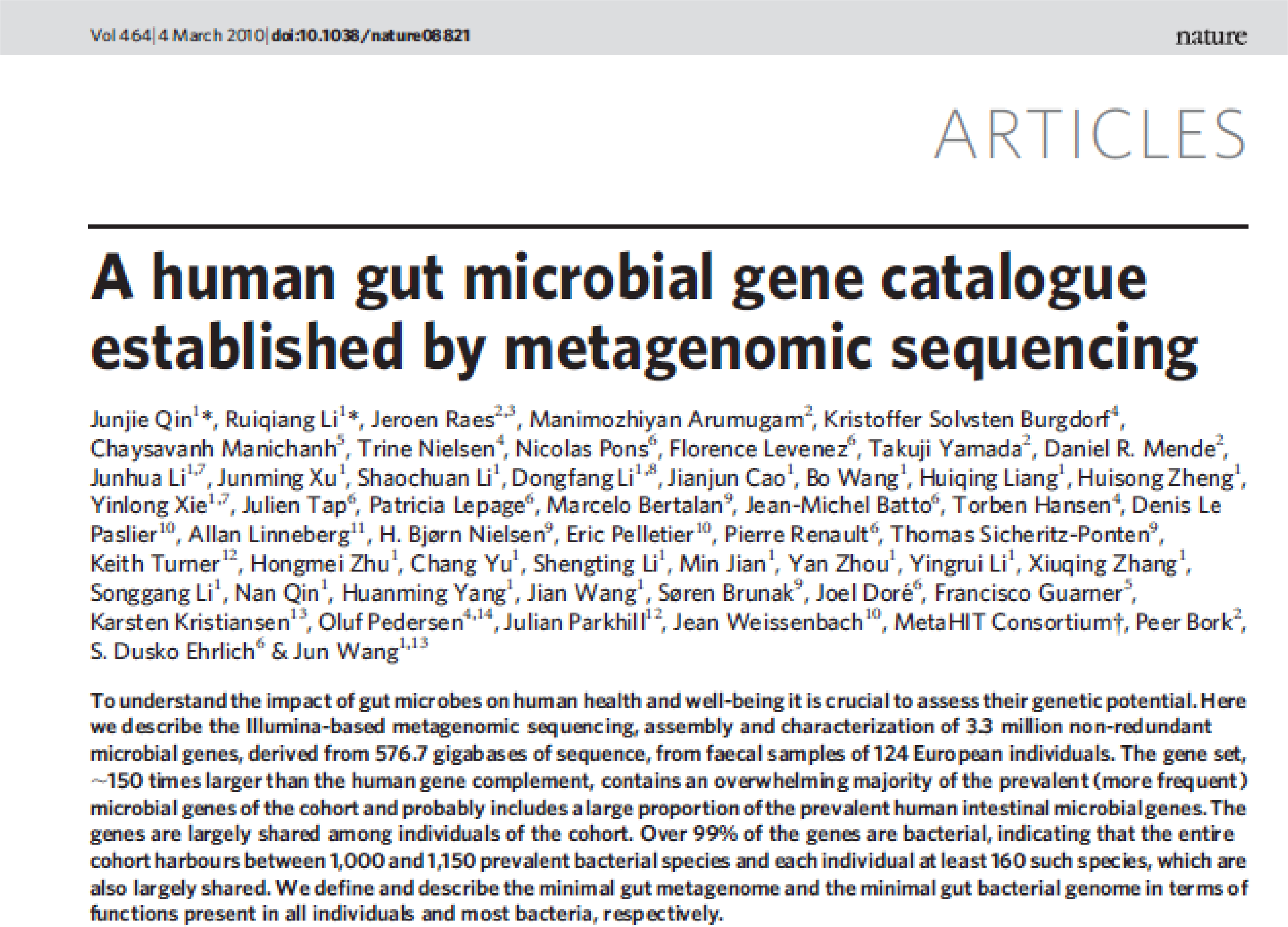
Meta-HIT project

The project objectives: association of bacterial genes with human health an disease

The central objective of our project is to establish associations between the genes of the human intestinal microbiota and our health and disease. We focus on two disorders of increasing importance in Europe, Inflammatory Bowel Disease (IBD) and obesity.

http://www.metahit.eu

MetaHIT paper



Vol 464| 4 March 2010| doi:10.1038/nature08821

nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

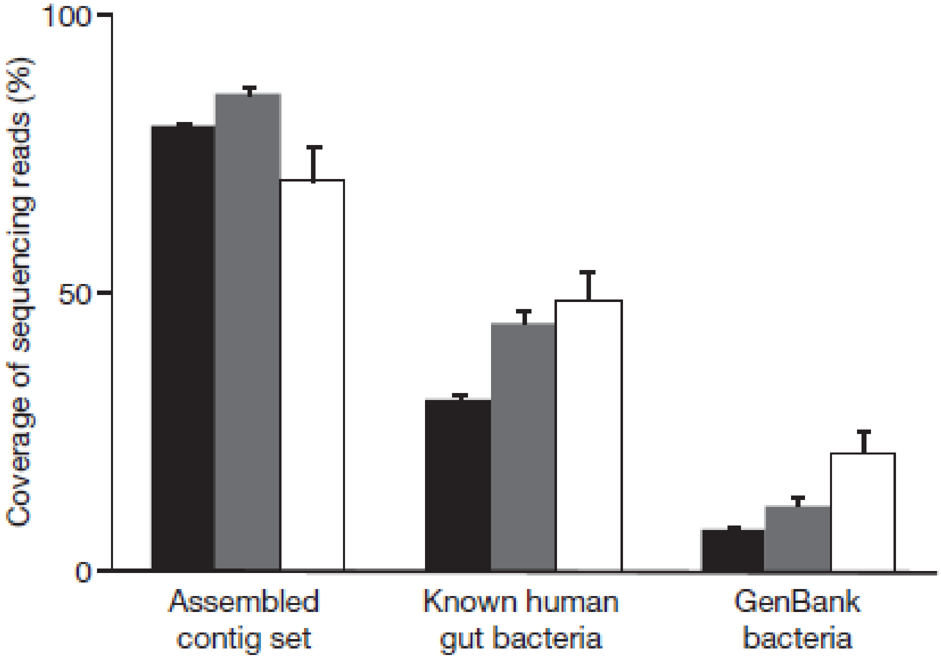
Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ∼150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

MetaHIT summary

- 8 billion reads
- 576Gb of sequence data
- 42% of reads assembled into 6.6 million contigs
- N50 contigs length of 2.2 kb

- 81% of genes un-annotated



More reference genomes are needed!

Meta HIT

Metagenomics of the Human Intestinal Tract

European research project

INTRODUCTION

Since 2008, researchers with the European consortium MetaHIT have been analyzing the collected genomes of the microorganisms present in our intestine : the microbiota.

Budget

22 million euros

The 4 year program was financed in large part by the European Union under the FP7 (*7th Framework Programme*).

Laboratories

8 countries

14 research & industrial

institutions are involved in the consortium, with more than 50 researchers and cooperation between Europe and China.

The microbiota

The microbiota is an ecosystem composed of billions of bacteria that make up a veritable "organ." Within 24 hours of birth, these bacteria colonize our digestive tract to form our intestinal microbiota (2kg for adults). MetaHIT focuses on the digestive tract since it is where the largest and most diversified bacterial community lives in our body.

RESEARCH

Little understood until now, the intestinal microbiota interests researchers as an avenue of inquiry to explain the evolution of chronic diseases.

Observations

↗ chronic diseases

↘ infectious diseases

Observations made in the past 50 years cannot be solely explained by variations of our genome.

Research themes

Nutrition. Better knowledge of the intestinal microbiota of individuals will enable the nutritional needs to adapt to everyones specific nutrient needs.

Medicine. With the study of the microbiota and the established catalogue of genes, we can have an unprecedented overview of the microbiota in healthy individuals and in patients. With the discovery of enterotypes we can imagine the upcoming development of new diagnostic or even prognostic tools for human health.

DEFINITION

***Enterotypes**

There are three in the world's population, each characterized by a predominant bacterial population.

FINDINGS

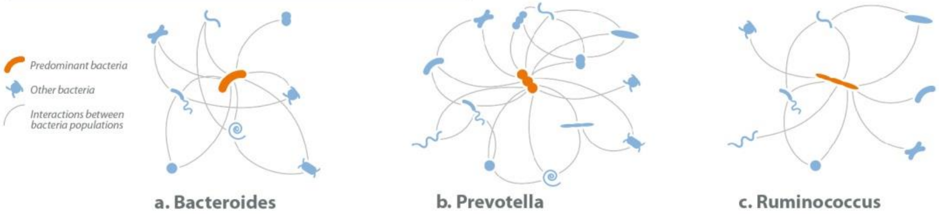
The MetaHIT consortium published two major findings in the scientific journal Nature : an established catalog of bacterial genes in the intestine; and the discovery of enterotypes.

Genome sequencing

3,3 million genes

The gut bacterial gene catalog, which can be compared to a *molecular scanner*, was established by metagenomic high throughtput sequencing and allows the observation of the human gut microbiome.

Discovery of the 3 enterotypes*



PERSPECTIVES

MetaHIT opens avenues for further efforts in the field of human microbiome research : early detection of chronic diseases, personalized medicine and more healthful food.

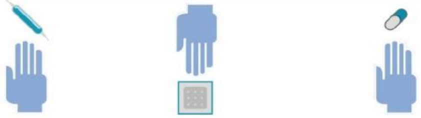
Chronic diseases

Disturbances in the microbiota can be early warning signs for certain diseases like Crohn's disease or diabetes.

Nutritional impact

If it is possible to reveal early warning signs of obesity, one can imagine nutritional intervention and diet advice being used to reestablish a healthy microbiota. The possibility of intervening directly in the flora, in the case of disturbance to the intestinal ecosystem, could also be envisioned.

Personalized medicine



Classification by enterotype will help in the development of diagnostic tools able to reveal cases where a planned treatment would not be effective, and to adapt it accordingly.

The gene set

Metagene prediction on the contigs:

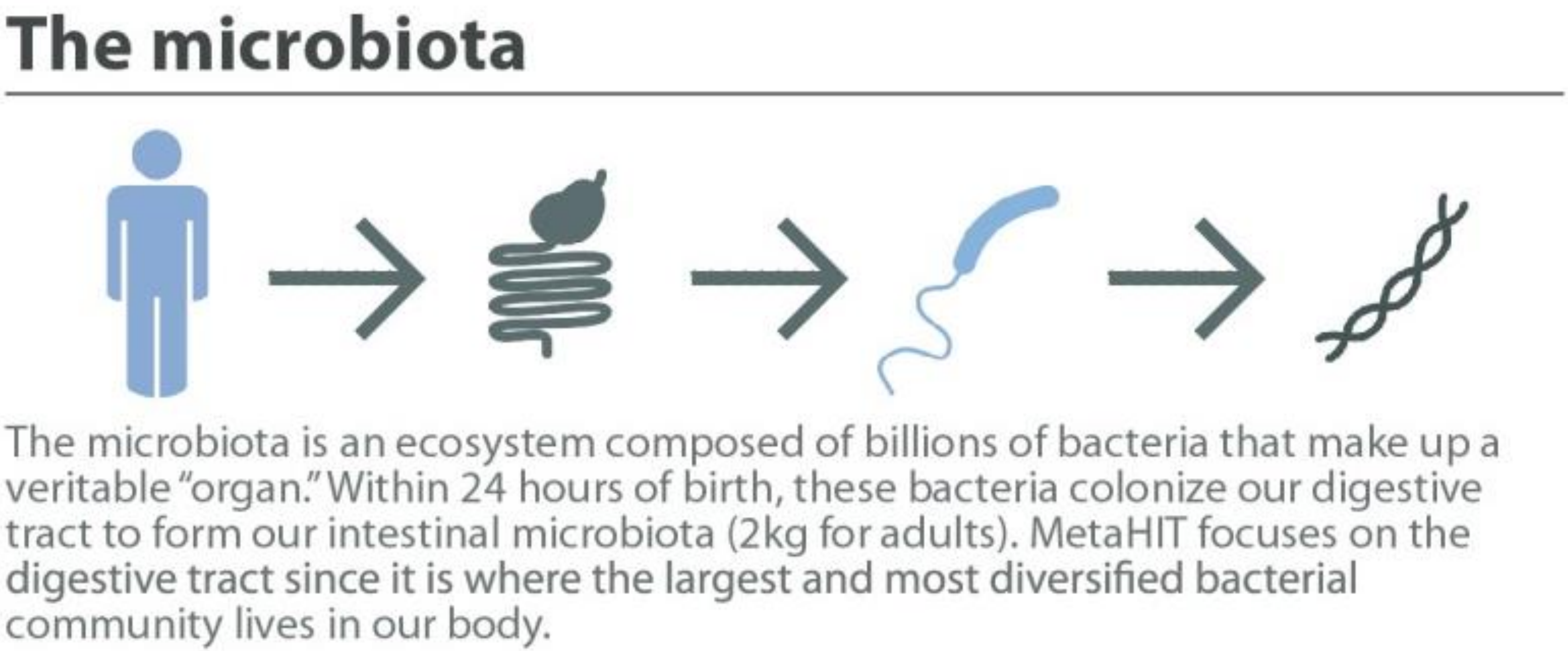
- 14 million ORFs >100 bp

Removal of redundancy : ≥ 95 % nucleotide identity, ≥ 90 % of the length of the shorter ORF

- 3.3 million ORFs, 150 times human gene complement

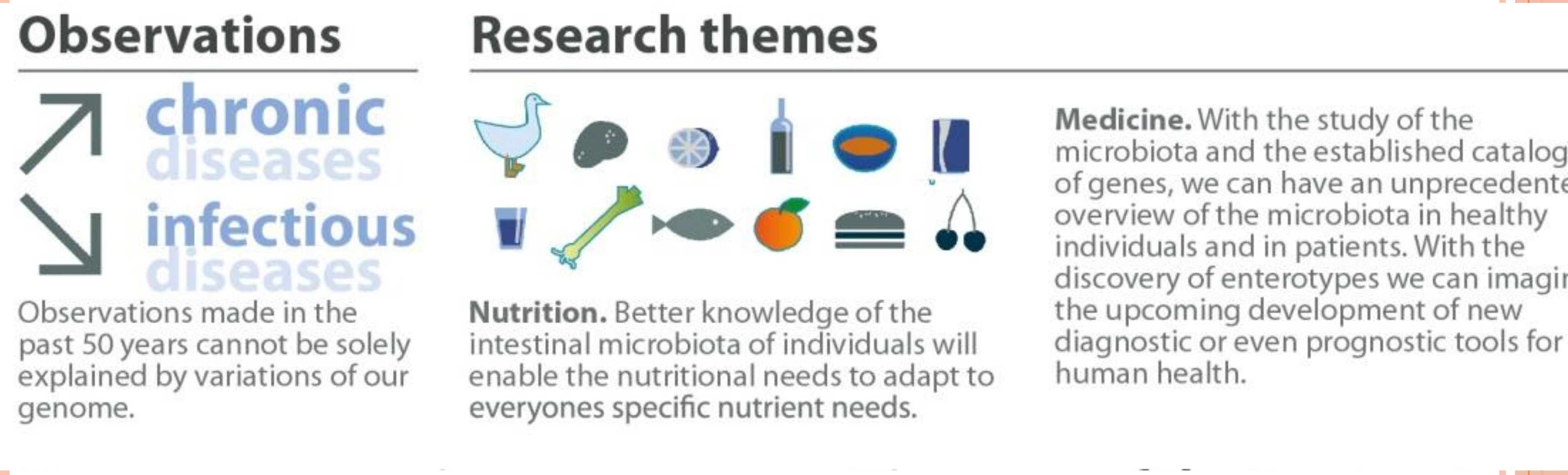
ORFs are identified if present at relative abundance

~$7\text{x}10^{-7}$; we name them “prevalent genes”



The microbiota

The microbiota is an ecosystem composed of billions of bacteria that make up a veritable "organ." Within 24 hours of birth, these bacteria colonize our digestive tract to form our intestinal microbiota (2kg for adults). MetaHIT focuses on the digestive tract since it is where the largest and most diversified bacterial community lives in our body.



Observations

chronic diseases

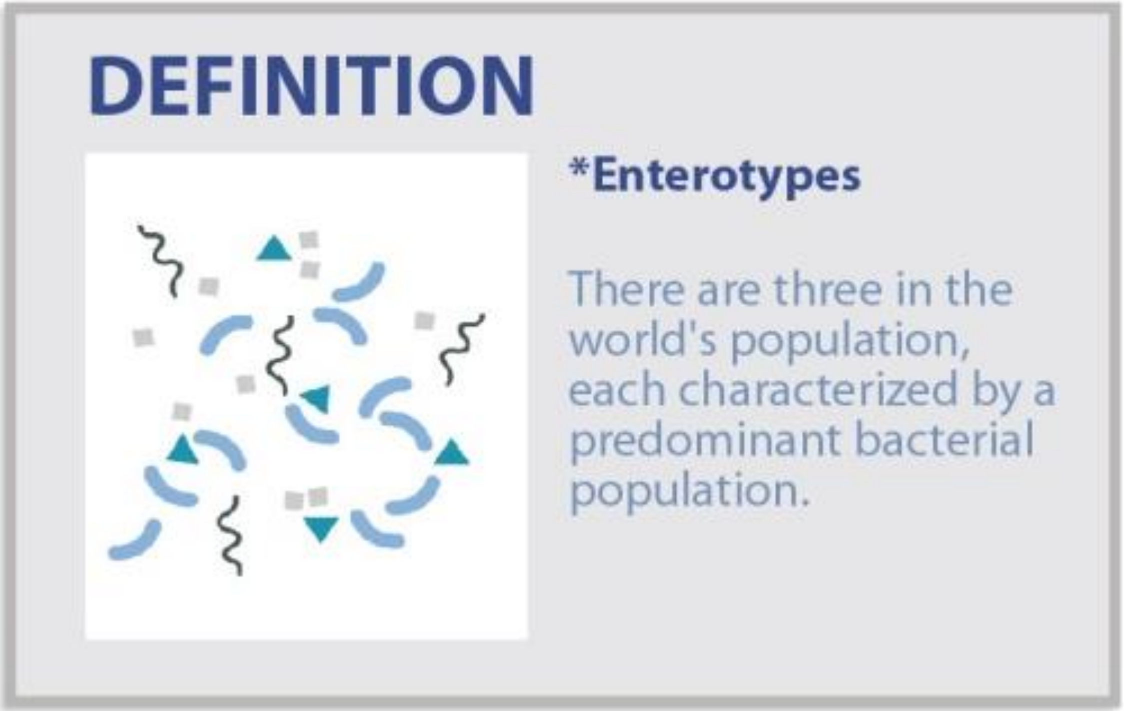
infectious diseases

Observations made in the past 50 years cannot be solely explained by variations of our genome.

Research themes

Nutrition. Better knowledge of the intestinal microbiota of individuals will enable the nutritional needs to adapt to everyones specific nutrient needs.

Medicine. With the study of the microbiota and the established catalog of genes, we can have an unprecedente overview of the microbiota in healthy individuals and in patients. With the discovery of enterotypes we can imagi the upcoming development of new diagnostic or even prognostic tools for human health.

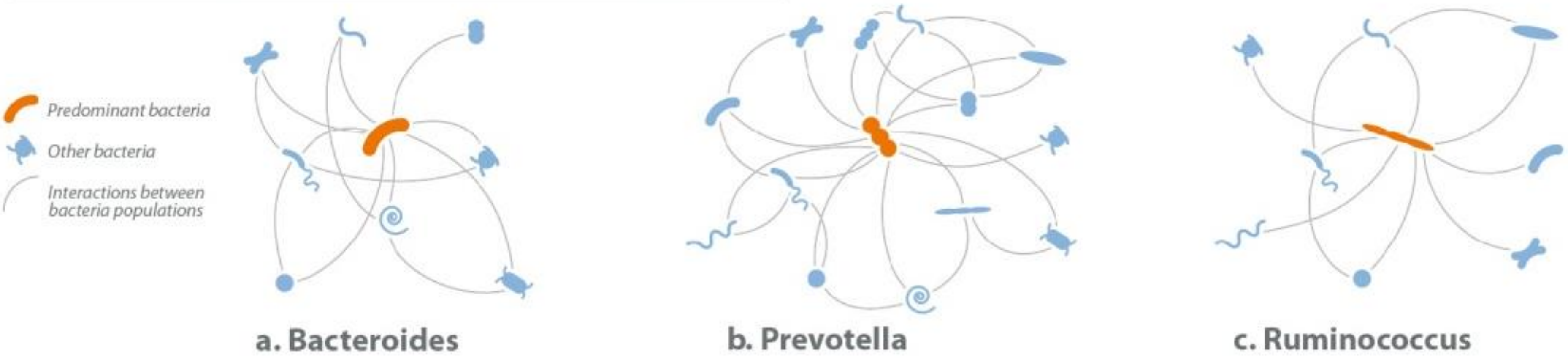


DEFINITION

***Enterotypes**

There are three in the world's population, each characterized by a predominant bacterial population.

Discovery of the 3 enterotypes*



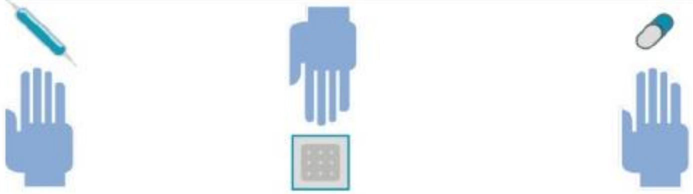
Chronic diseases

Disturbances in the microbiota can be early warning signs for certain diseases like Crohn's disease or diabetes.

Nutritional impact

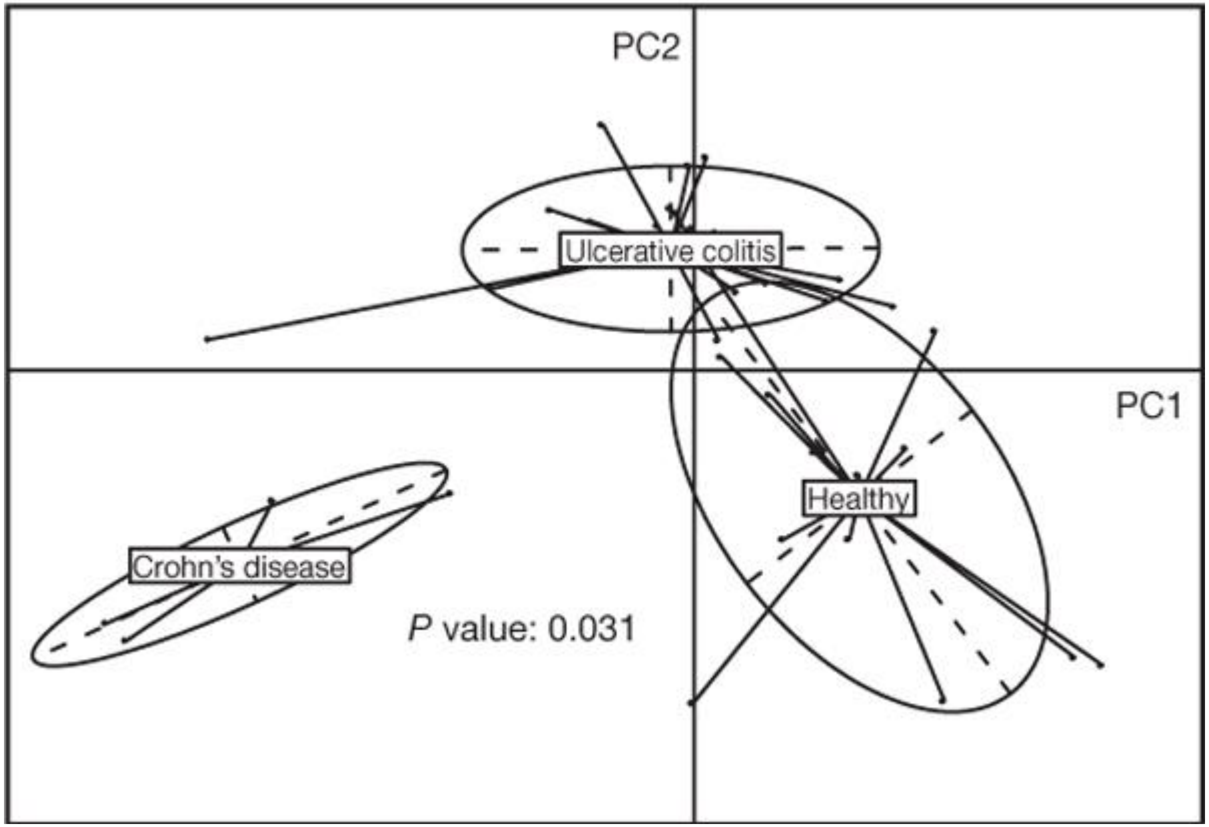
If it is possible to reveal early warning signs of obesity, one can imagine nutritional intervention and diet advice being used to reestablish a healthy microbiota. The possibility of intervening directly in the flora, in the case of disturbance to the intestinal ecosystem, could also be envisioned.

Personalized medicine



Classification by enterotype will help in the development of diagnostic tools able to reveal cases where a planned treatment would not be effective, and to adapt it accordingly.

PCA of 155 most abundant bacterial species in IBD patients and healthy controls (n=39)



IBD=inflammatory bowel disease

A human gut microbial gene catalogue established by metagenomic sequencing, Nature ***464****, 59-65(4 March 2010)*

Overview

- **What is metagenomics?**
  - Why?
  - Case study
  - **Assembly, ORFs and Gene finding**
  - Annotation

Metagenomic assemblies

- Much harder than single-genome assembly
  - Many identical or nearly identical reads
  - Reduce size by clustering data first at 100% identity
  - Cannot remove near-identical low abundance kmers to reduce memory requirements
    - These may be sequencing errors
    - Or may be sequences from low abundance organisms
  - Can try to focus on gene regions by identifying putative open reading frame start sites and start assembly there

- Still very early days. Hardware requirements large.

- Meta-Velvet
- Soapdenovo
- Euler

Ye Y, Tang, H. An orfome assembly approach to metagenomics 2009 J. Bioinform Comput Biol 7: 455-471

Gene calling metagenomic assemblies

Gene calling

- Finding open reading frames (ORFs) is challenging when assemblies of gene may only be partial
- Start and/or stop coding may be missing
- Traditional HMM-based methods (e.g. Genemark) fail
- However, simulations have shown that 85-90% of genes can be accurately called – although this is best case scenario

- Gene families coding for proteins are expected to be under selective pressure
- One method is to select all reading frames from any ORF identified and use only those which appear to be under selective pressure
- This may miss ORFs under less selective pressure

Mavromatis et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methos. 2007. Nat Methods 4:495-500

Yooseph, et al. Gene identification and classification in microbial metagenomic sequence data via incremental clustering 2008. BMC Bioinformatics 9:182

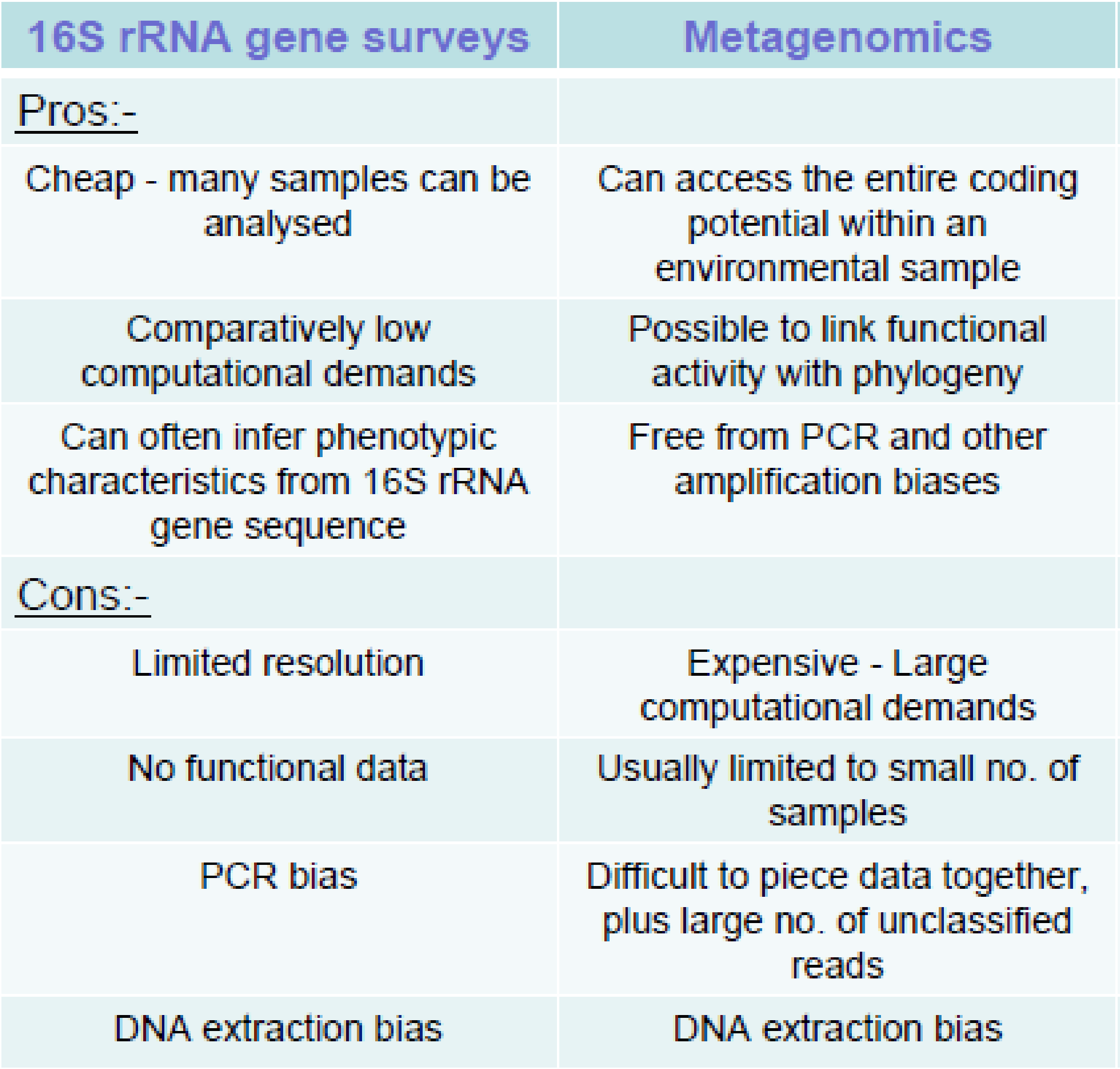
But…

Many organisms and genes are still unknown to science

Therefore homology-based annotation and even motif and HMM based annotation will only provide reliable annotation for those proteins we already know about

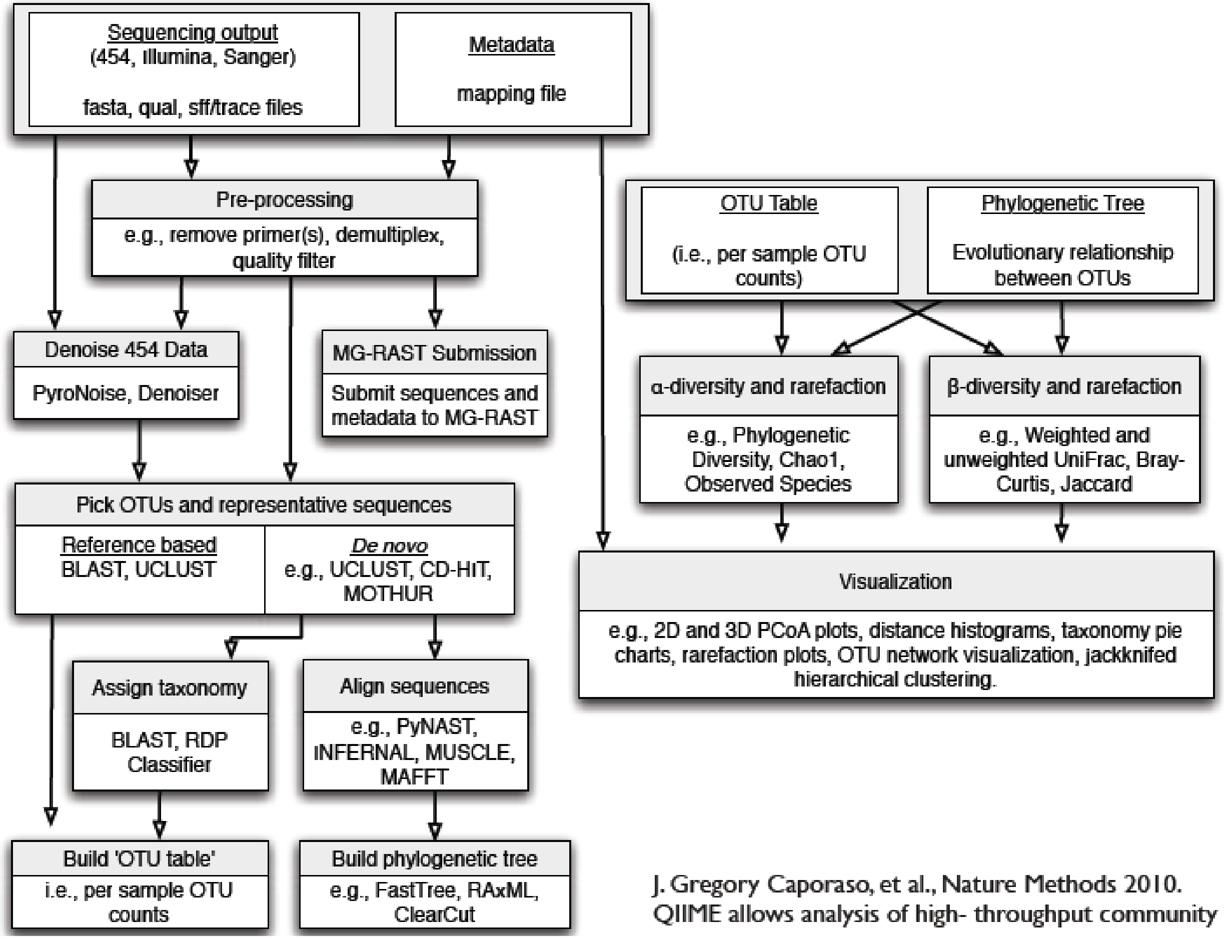
Current methods will still miss known genes

Summary



16S rRNA gene surveys	Metagenomics
Pros:-	
Cheap - many samples can be analysed	Can access the entire coding potential within an environmental sample
Comparatively low computational demands	Possible to link functional activity with phylogeny
Can often infer phenotypic characteristics from 16S rRNA gene sequence	Free from PCR and other amplification biases
Cons:-	
Limited resolution	Expensive - Large computational demands
No functional data	Usually limited to small no. of samples
PCR bias	Difficult to piece data together, plus large no. of unclassified reads
DNA extraction bias	DNA extraction bias

QIIME – Quantitative Insights Into Microbial Ecology



Sequencing output
(454, Illumina, Sanger)
fasta, qual, sff/trace files

Metadata
mapping file

Pre-processing
e.g., remove primer(s), demultiplex, quality filter

Denoise 454 Data
PyroNoise, Denoiser

MG-RAST Submission
Submit sequences and metadata to MG-RAST

Pick OTUs and representative sequences

Reference based
BLAST, UCLUST

De novo
e.g., UCLUST, CD-HIT, MOTHUR

Assign taxonomy
BLAST, RDP Classifier

Align sequences
e.g., PyNAST, INFERNAL, MUSCLE, MAFFT

Build 'OTU table'
i.e., per sample OTU counts

Build phylogenetic tree
e.g., FastTree, RAxML, ClearCut

OTU Table
(i.e., per sample OTU counts)

Phylogenetic Tree
Evolutionary relationship between OTUs

α-diversity and rarefaction
e.g., Phylogenetic Diversity, Chao1, Observed Species

β-diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray-Curtis, Jaccard

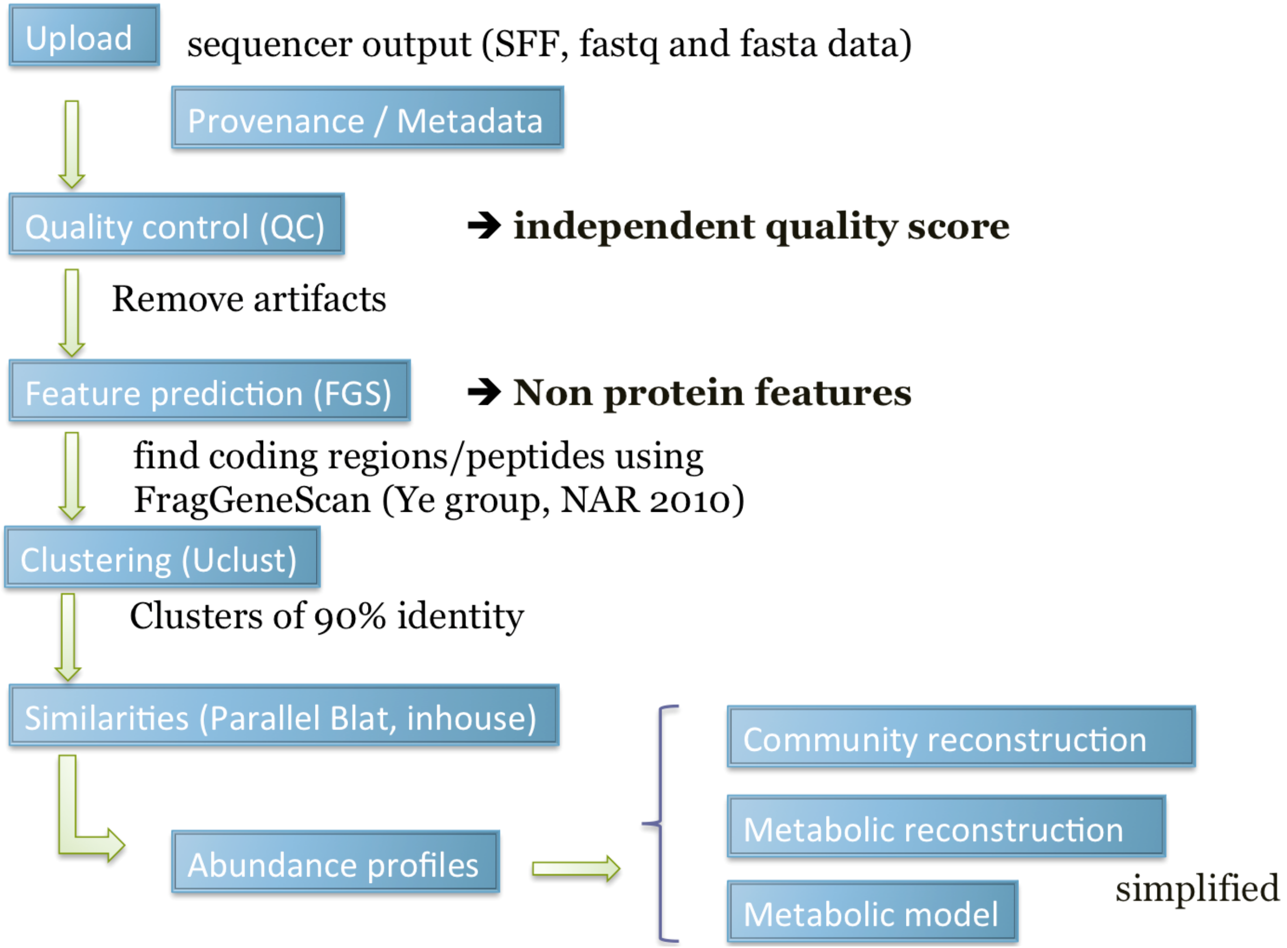
Visualization
e.g., 2D and 3D PCoA plots, distance histograms, taxonomy pie charts, rarefaction plots, OTU network visualization, jackknifed hierarchical clustering.

J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

The MG-RAST pipelines

MG-RAST has a number of pipelines with some user adjustable parameters. These fully automated pipelines create data sets that allow comparison between multiple data sets.

The following figure gives a simplified overview of the various steps in our pipeline.



Upload

sequencer output (SFF, fastq and fasta data)

Provenance / Metadata

Quality control (QC)

➔ **independent quality score**

Remove artifacts

Feature prediction (FGS)

➔ **Non protein features**

find coding regions/peptides using
FragGeneScan (Ye group, NAR 2010)

Clustering (Uclust)

Clusters of 90% identity

Similarities (Parallel Blat, inhouse)

Abundance profiles

Community reconstruction

Metabolic reconstruction

Metabolic model

simplified