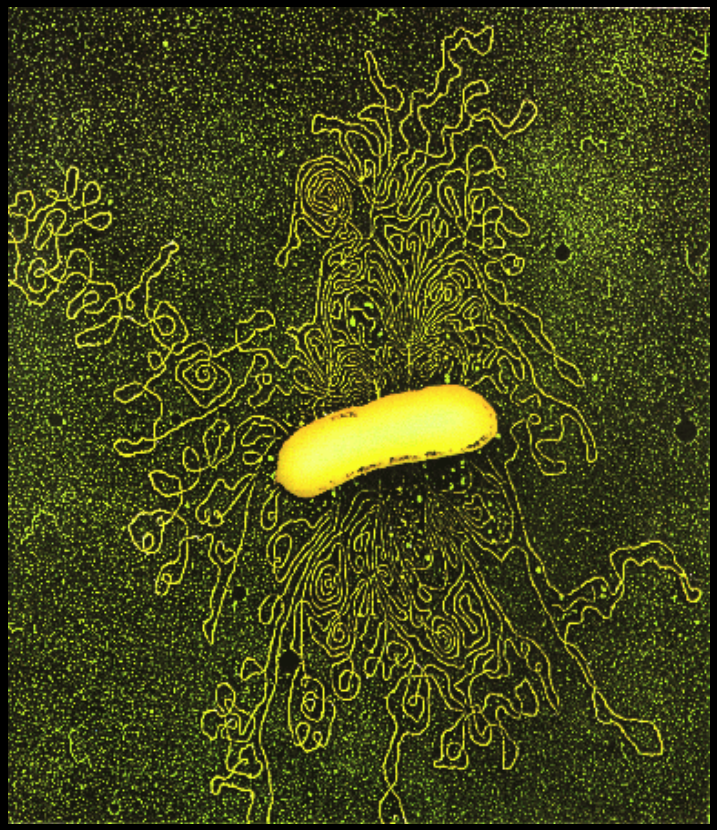
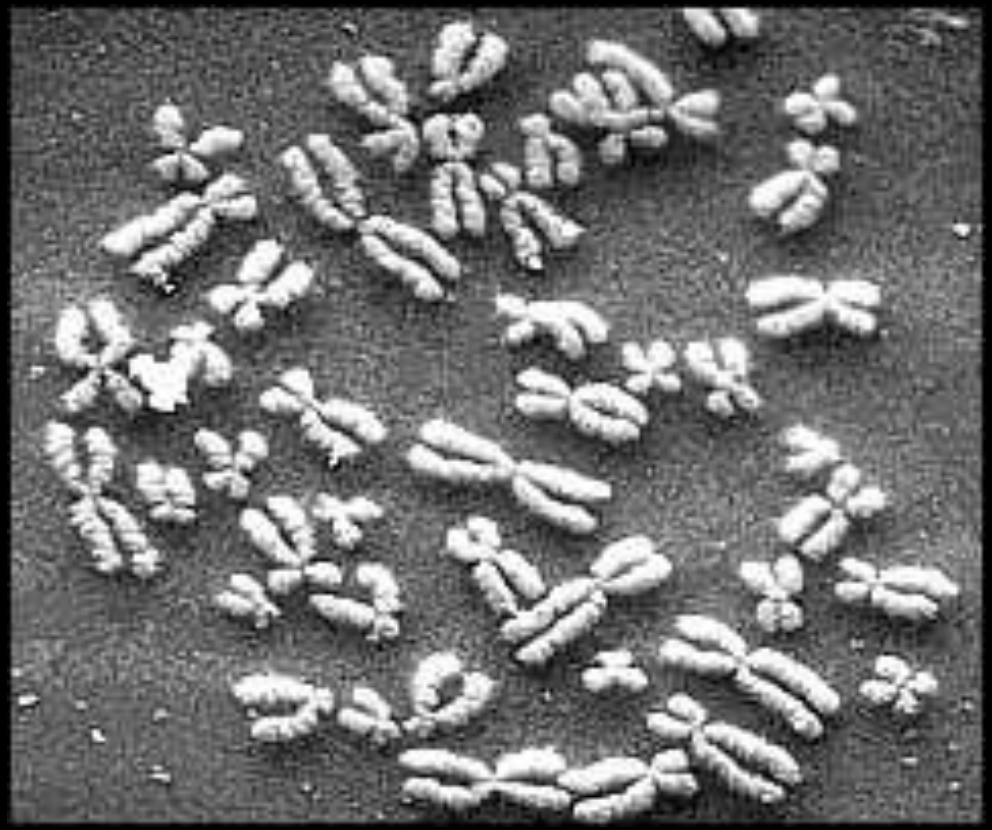
Environmental Microbiology

„Poly-omics I“ (VO 10)





Bettina Siebers

Microbe of the Day

Archaea

- Phylum: Euryarchaeota (7 Classes)
  - Class: Thermoplasmata (1 order)
    - Order: Thermoplasmatales (3 families)
      - Family: Picrophilaceae (1 genus)
        - » Genus: *Picrophilus* (2 species)
          - species: *Picrophilus torridus*
          - species: *Picrophilus oshimae*
      - Family: Thermoplasmataceae
      - Family: Ferroplasmataceae

Thermoplasmales

- ***Thermoplasma, Ferroplasma***, and ***Picrophilus*** are **extremely acidophilic thermophiles** that form their own phylogenetic family of *Archaea* inhabiting **coal refuse piles** and **highly acidic solfataras**.
- Cells of *Thermoplasma* and *Ferroplasma* **lack cell walls** and thus resemble the mycoplasmas in this regard.
- **Picrophilus** has an **S-layer**

Archaea: *Picrophilus torridus*

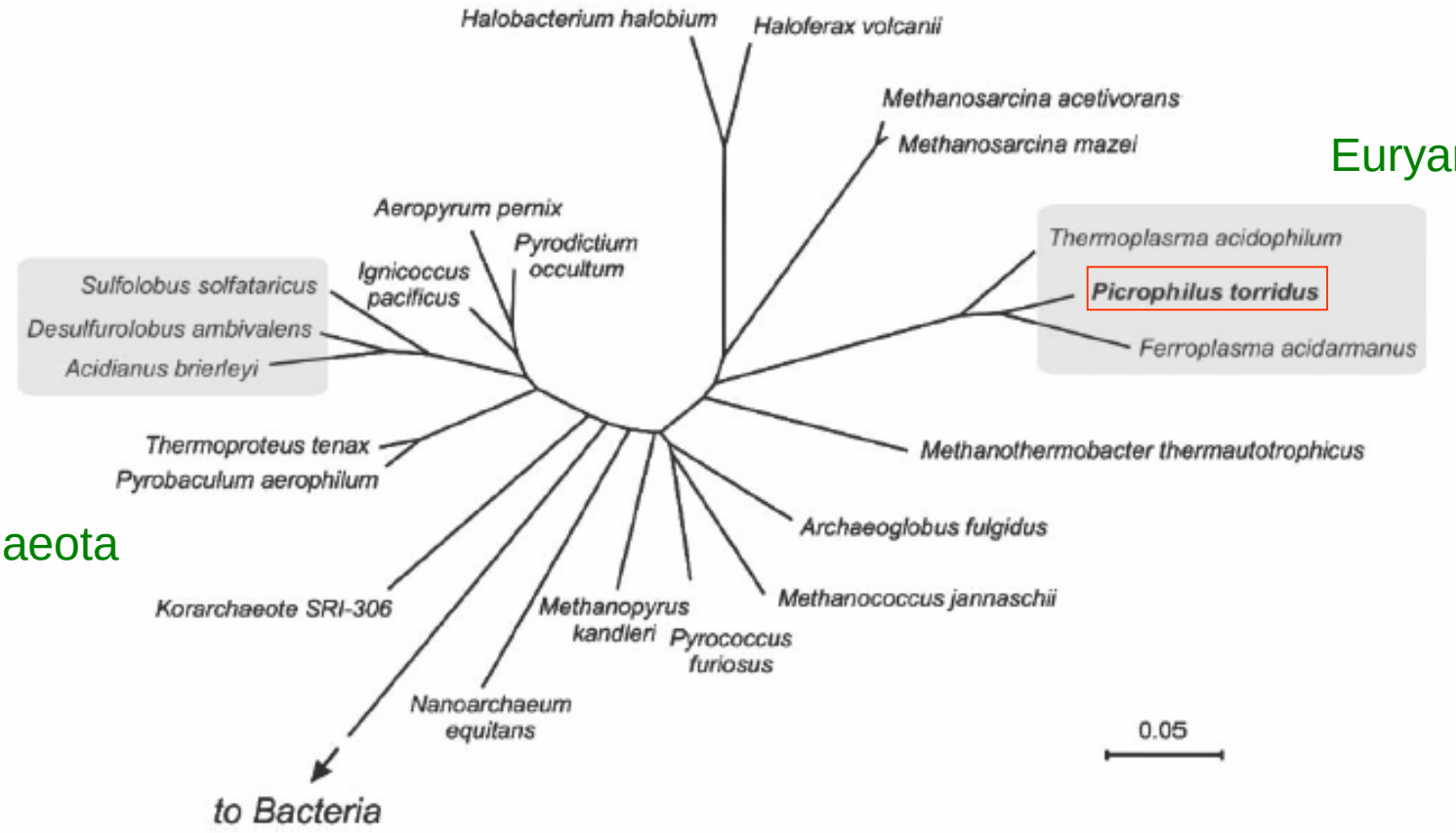
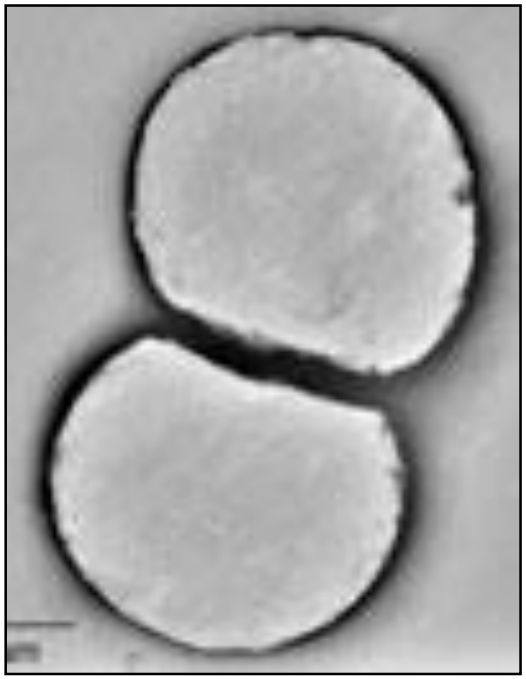


Fig. 1. 16S rRNA phylogenetic tree. Highlighted are the two thermoacidophilic groups of the archaea. Sequences were aligned with the CLUSTALW algorithm. The tree was build by neighbor joining by using the Kimura 2-parameter for distance calculation.

P. torridus is found within the ORDER Thermoplasmatales which is made up of thermoacidophilic organisms of the euryarchaeal branch of the DOMAIN *Archaea*

Archaea: *Picrophilus torridus*

Archaea; Euryarchaeota; Thermoplasmata; Thermoplasmatales; Picrophilaceae; Picrophilus
Domain, Phylum, Class, Order, Family, Genus



Found in northern Japan in soil heated by underground volcanic activity.

The word *torridus* roughly means “burnt and dried.”

- It is **thermoacidophilic**: lives at temperatures around **60°C at pH 0 (pH optimum pH 0.7).**
- Maintains intracellular pH 4.6.
  - Other thermoacidophilic organisms are able to maintain a near neutral intracellular pH.
- Cells (with S-layer) have a **specially adapted membrane** with very low proton permeability. It becomes unstable at pH 7.

Fütterer *et al*., 2004, PNAS 101:9091-9096

Analyses of Microbial Communities

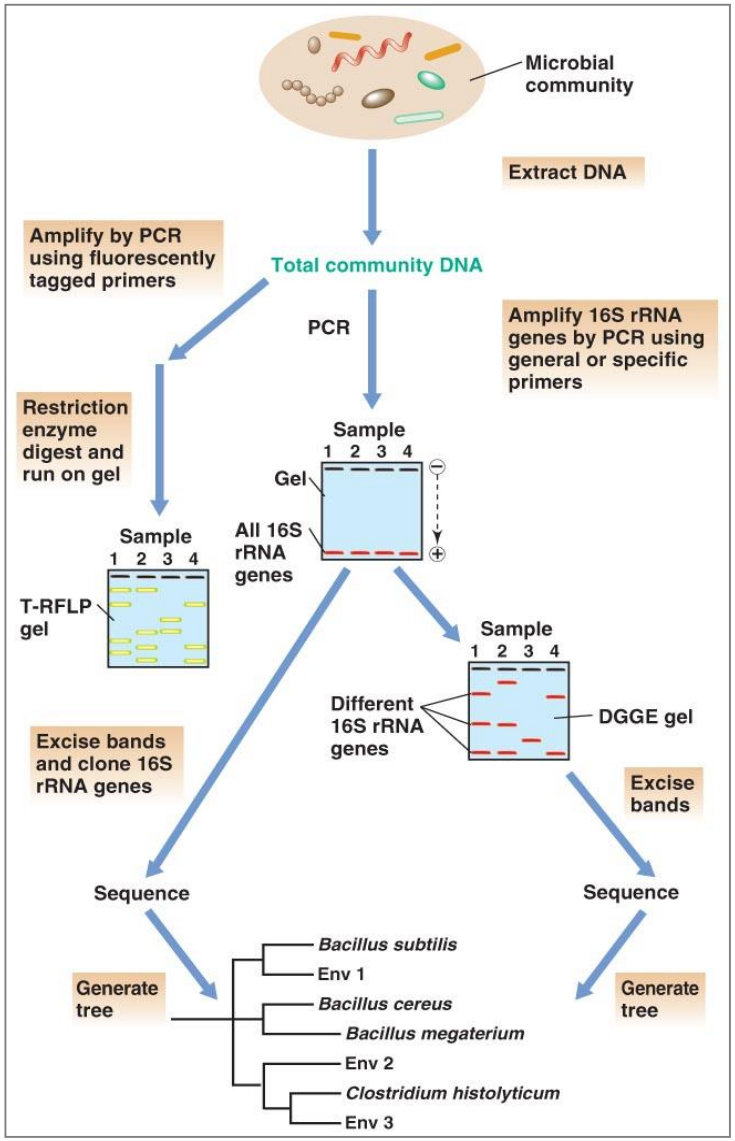
- **I. Culture-dependent methods**
  - Enrichment and Isolation
- **II. Molecular (Culture-Independent)**
  - (A) Viability and Quantification Using **Staining Techniques**
  - (B) **Genetic stains** (FISH, chromosome painting, ISRT FISH, CARD FISH)
  - (C) Linking specific genes to specific organisms **using PCR (clone library, DGGE, T-RFLP)**
  - (D) Environmental Genomics (**Metagenomics**)
- **III. Measuring Microbial Activities in Nature**
  - Radioisotopes (Fish-MAR), microelectrodes, stable isotopes, NanoSIMS

II Molecular (Culture-Independent) Analyses of Microbial Communities

"Molecular Microbial Ecology"

IIC Linking specific genes to specific organisms using PCR

Linking specific genes to specific organisms using **PCR**



Steps in Single Gene Biodiversity

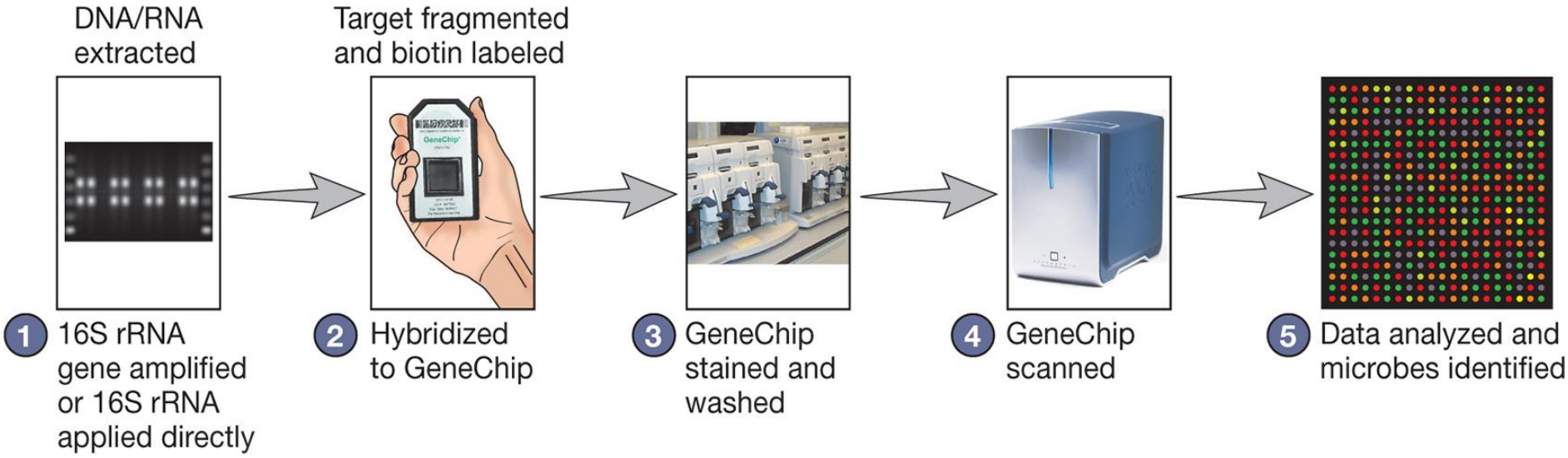
16S rDNA based methods

Averaging methods (provide an **overview of diversity** but say nothing about the spatial arrangement in the original sample)

- **Clone library**
- **Denaturing gradient gel electrophoresis (DGGE)**
- **Terminal Restriction Fragment Length Polymorphism (T-RFLP)**

Phylochip

- Phylogenetic oligonucleotide arrays (**POAs**)
- Community genome arrays (**CGAs**)
  - SSU rRNA probes along with probes for certain genes
  - allows for measuring microbial activity



II Molecular (Culture-Independent) Analyses of Microbial Communities

"Molecular Microbial Ecology"

IID Environmental Genomics

"Metagenomics"

Not PCR-dependent !

Single Gene vs Environmental Genomics

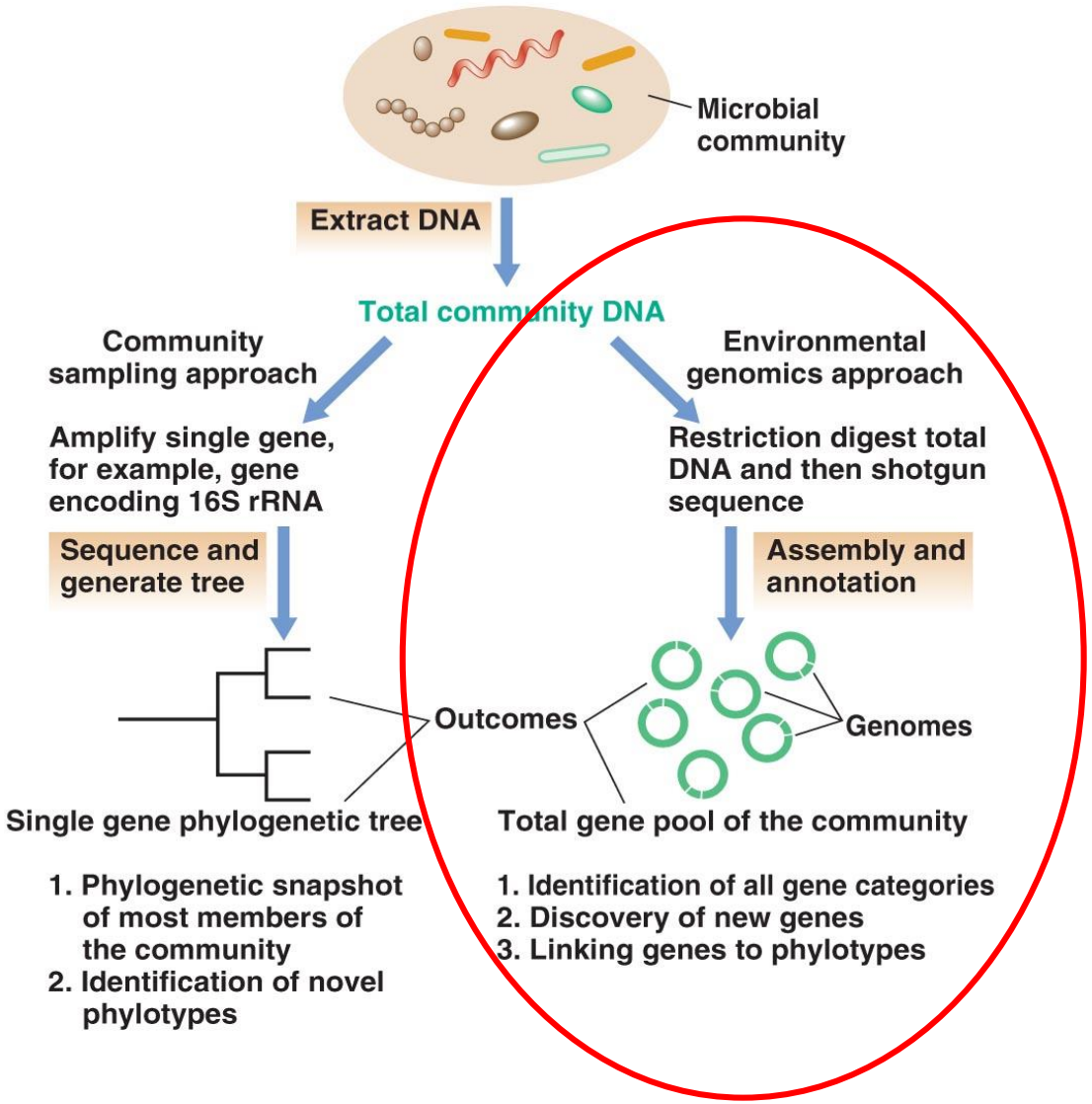


Figure 22.16

Molecular community analysis
-Biodiversity of a single gene
(PCR based 16S/18S RNA)

versus

Metagenomics
-**All genes** in the microbial community
„**metagenome**“

Metagenomics

„Who is there and what are they doing ?“

- **Metagenomics:** All genes in the microbial community „metagenome“
- Aim: Detect **as many genes as possible** and determine to which **phylogenetic „scaffold“** they belong.
- Done by sequencing overlaps to the genes that include phylogenetic markers (e.g. 16S rRNA genes)
- Much **higher costs** but more in-depth
- **Avoids „selectivity“ of PCR** (primers), all genes are sequenced whether they are amplifiable or not

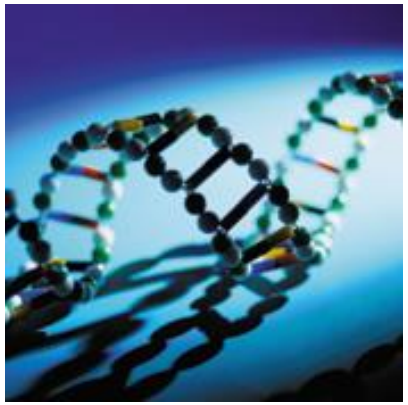
Great Potential of Metagenomics

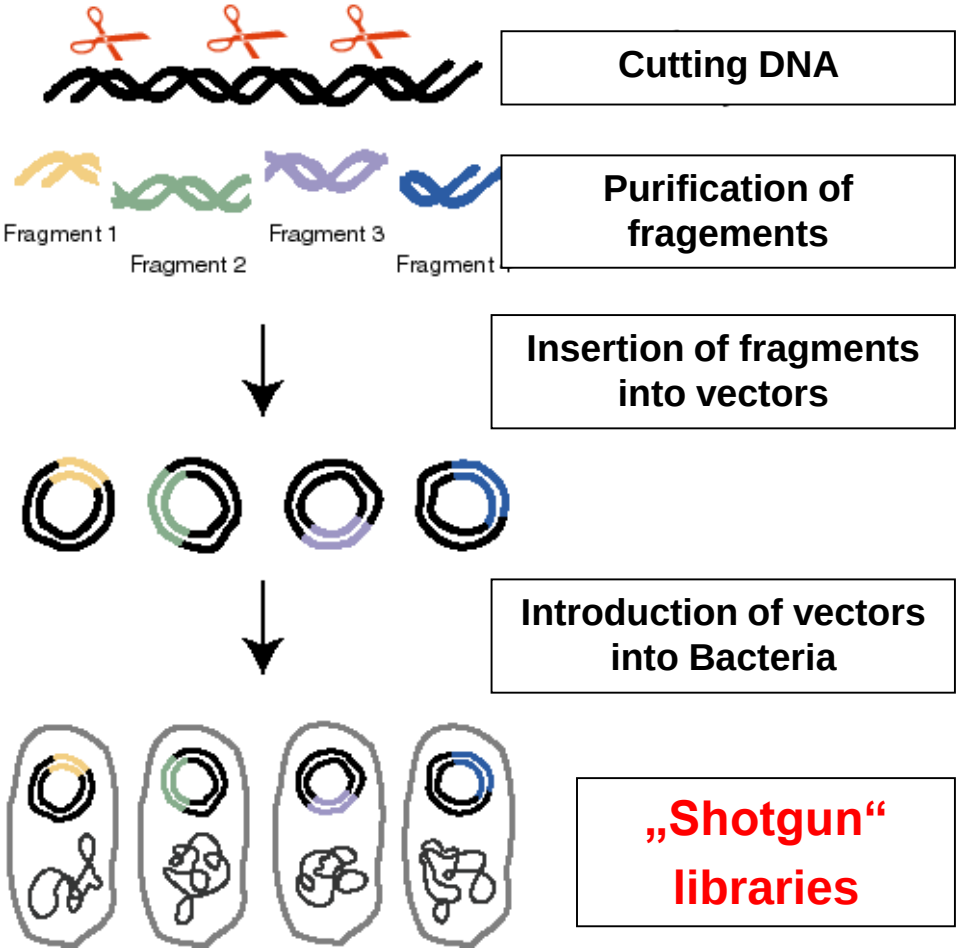
- Detects **new genes in known organisms** and **known genes in new organismsm**
- E.g. genes encoding **ammonia monooxygenase** „key enzymes of ammonia-oxidizing Bacteria“ in **Archaea** (These Archaea have never been described!)
- Functional metagenomics, e.g. **screening of new catalysts**

Table 29.4 Example Genes Used to Assess Community Function and Metabolic Diversity

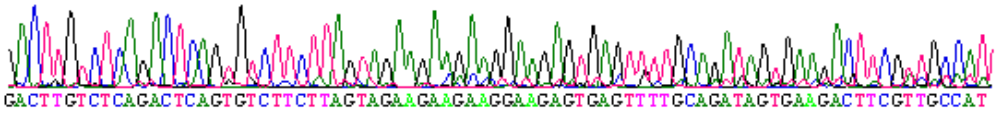
Process	Gene	Gene Product
Ammonium oxidation	*amoC, amoC, amoB*	Ammonia monooxygenase
Anammox	*hzf*	Hydrazine hydrolase
Dissimilatory nitrate reduction	*narG* *nosZ* *nir*	Nitrate reductase Nitrous oxide reductase Nitrite reductase
Nitrogen fixation	*nifH, nifD, nifK*	Different subunits of nitrogenase
Methanogensis	*mcrA*	Methyl coenzyme M reductase
Methane oxdition	*pmoA*	Methane monooxygenase
Sulfate reduction	*apsA*	Adenosine phosphosulfate reductase

Classical „Shotgun Sequencing“ Approach





„Shotgun Sequencing“



Shotgun sequencing of libraries

Assembley

16S rRNA gene

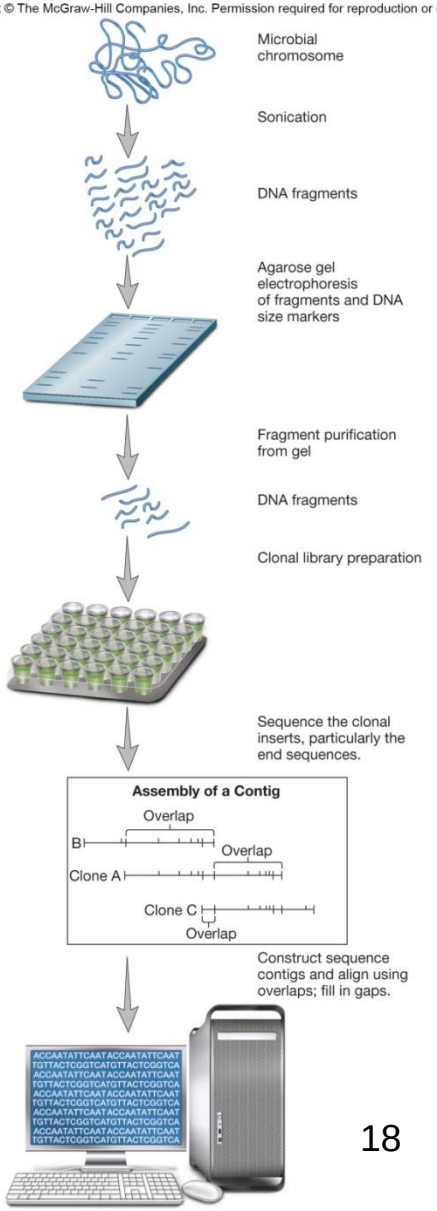
Contig

Contig

Phylogenetic marker !!

Whole-Genome Shotgun Sequencing

- Developed in 1995 by J. Craig Venter and Hamilton Smith
- Four stage process
  - library construction
    - generates clones of portions of genome
  - random sequencing
    - determines sequences of clones
  - fragment alignment and gap closure
  - editing



Cloning Vectors and Creating Recombinant DNA

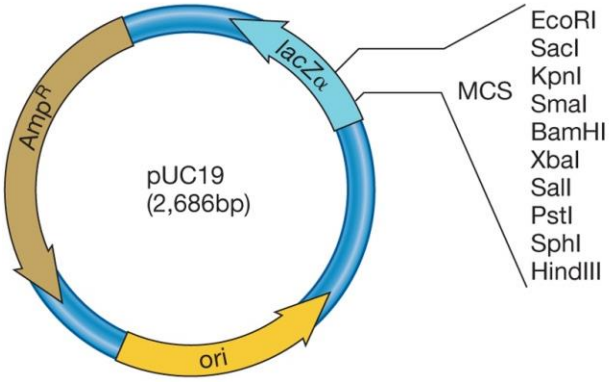
- There are four types of cloning vectors
  - plasmids (most commonly used)
  - phages and viruses
  - cosmids
  - artificial chromosomes

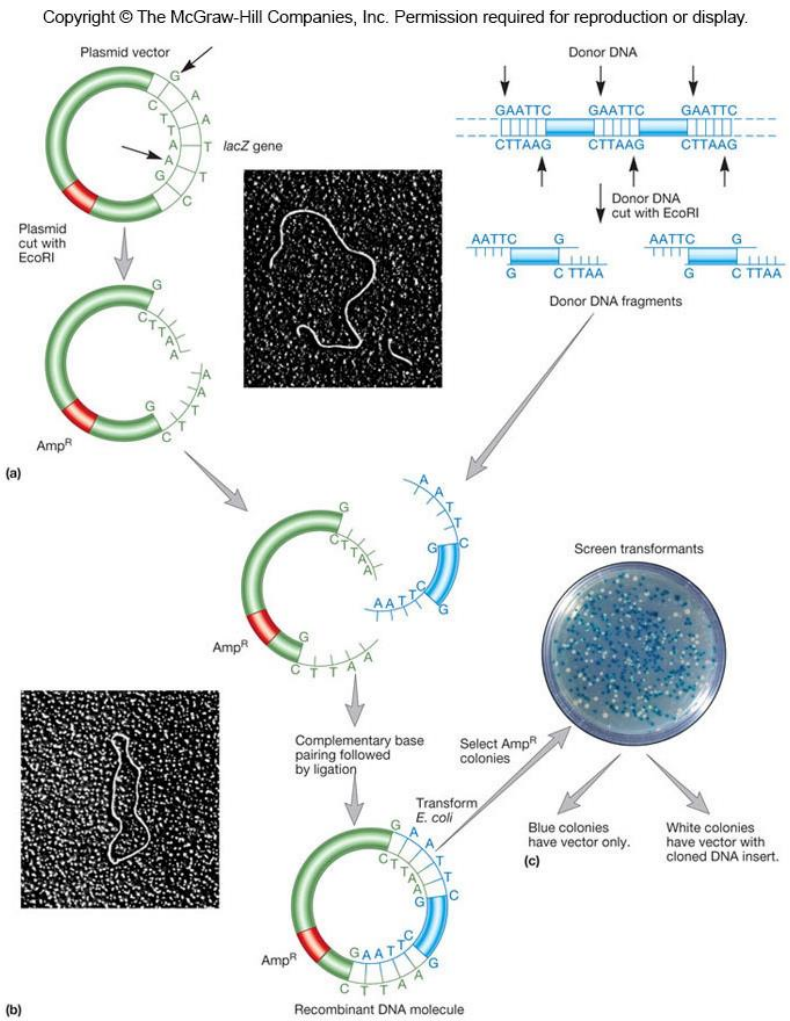
Table 17.3 Recombinant DNA Cloning Vectors

Vector	Insert Size (kb, 1 kb = 1,000 bp)	Example	Characteristics
Plasmid	<20 kb	pBR322, pUC19	Replicates independently of microbial chromosome so many copies may be maintained in a single cell
Bacteriophage	9–25 kb	λ 1059, λ gt11, M13mp18, EMBL3	Packaged into lambda phage particles; single-stranded DNA viruses such as M13 have been modified (e.g., M13mp18) to generate either double- or single-stranded DNA in the host
Cosmids	30–47 kb	pJC720, pSupercos	Can be packaged into lambda phage particles for efficient introduction into bacteria, then replicates as a plasmid
BACs (bacterial artificial chromosomes)	75–300 kb	pBAC108L	Modified F plasmid that can carry large DNA inserts; very stable within the cell
YACs (yeast artificial chromosomes)	100–1,000 kb	pYAC	Can carry largest DNA inserts; replicates in *Saccharomyces cerevisiae*

Plasmids

- Replicate autonomously and easy to purify
- Requirements for vectors
  - an origin of replication
  - a selectable marker
  - a multicloning site or polylinker
    - site that allows gene to be cloned to placed into the plasmid vector





Phage Vectors

- Are **engineered phage genomes** previously genetically modified to include **restriction sites**
- After **insertion of foreign DNA** the recombinant phage genome is packaged into the capsid and used to infect host cells "**tranduction**"

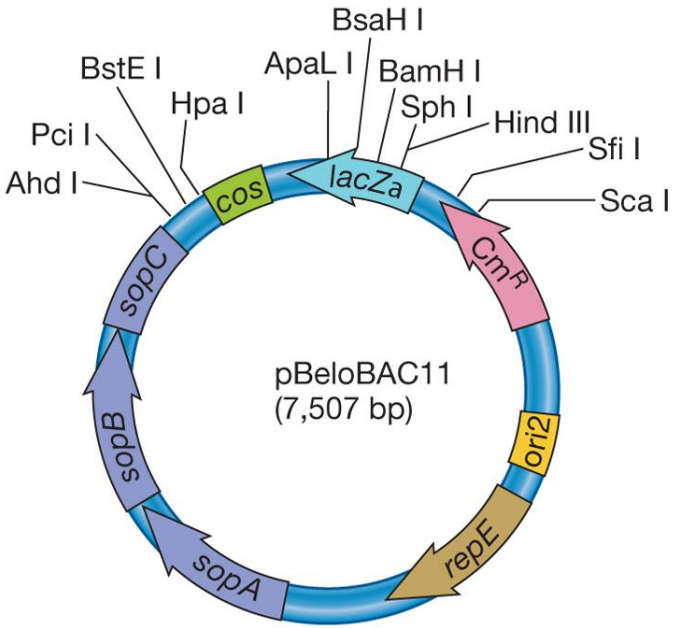
Cosmids

- Do not exist in nature "**artifical**"
- These vectors have been **constructed** to contain features from both **phages and plasmids**
  - they have a selectable marker, multiple cloning sites from plasmids, and a *cos (cohesive end (or sticky))* site from λ phage
  - phage then introduces recombinant DNA into *E. coli*
- Hybrid vectors **replicate as plasmids** but due to ***cos*** site can be also packed into phage capsids "**transduction**"

Artificial Chromosomes

(a) Yeast artificial chromosome (YAC)

TEL, yeast telomere at each end; CEN, centromere; ARS, yeast origin of replication; URA3, selectable marker; MCS, multiple cloning site



(b) Bacterial artificial chromosome (BAC)

repE, ensures formation of replication complex; *sopA,B,C* ensures proper partiioning of one newly replicated BAC to daughter cells; MCS; *lacZ*, blue/white colony screening; Cm^R, chloramphenicol (antibiotic) resistance

- Used when large fragments of DNA must be cloned
- **Yeast artificial chromosomes (YACs)**
  - may be unstable, for very large DNA fragments
- **Bacterial artificial chromosomes (BACs)**
  - Based on *E. coli* F fertility factor
  - played important role in the human genome project

GNN Genome News Network

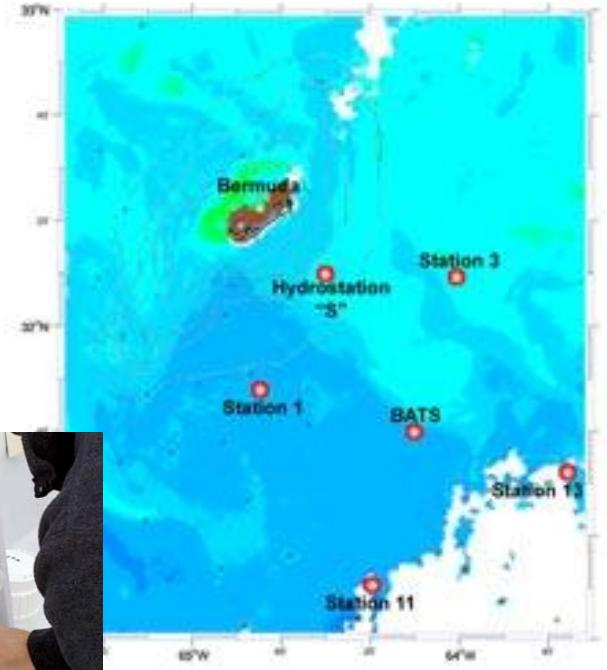
Home | About | Topics Search GNN

Exploring the Sargasso Sea

Scientists Discover One Million New Genes in Ocean Microbes

RV Weatherbird II

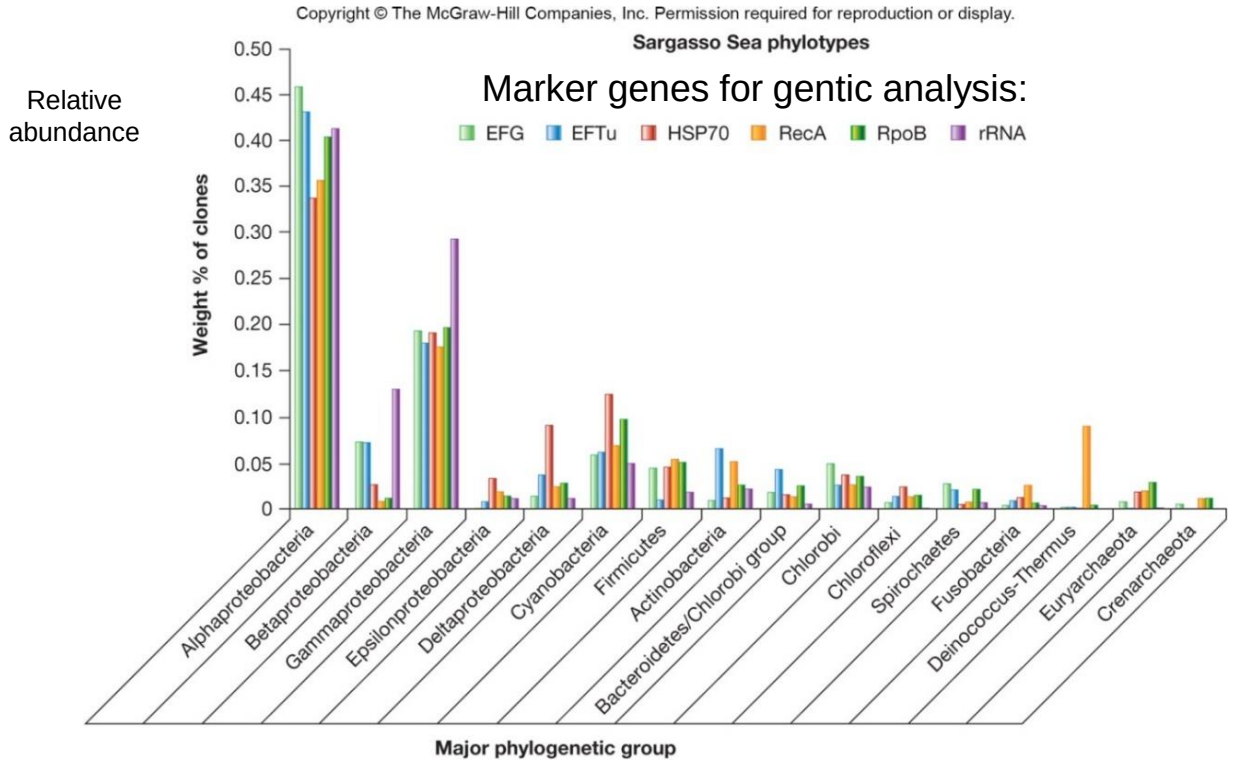
Craig Venter



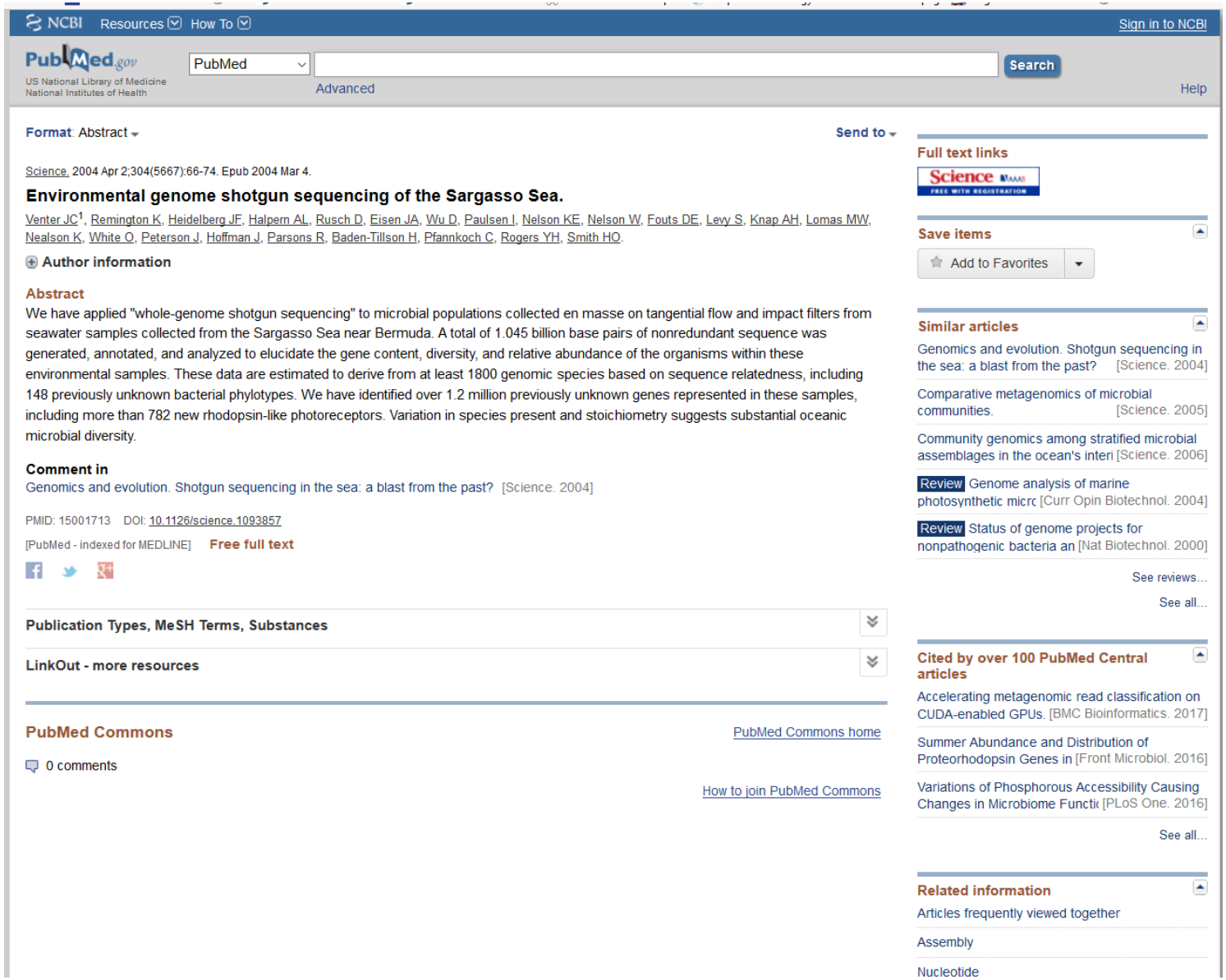
http://www.genomenewsnetwork.org/articles/2004/03/04/sargasso.php

Metagenomics - Phylotypes

- Genes that indicate taxonomy may aid in assessing metabolic activities of microbial communities
  - prokaryotic biodiversity of **Sargasso sea** (Atlantic Ocean, Bermuda) determined by **Craig Venter, Hamilton Smith et al.**; filtered sea water, extracted DNA, prepared genomic library and sequenced 1 billion base pairs; 1,800 phylotypes (145 unknown), discovery of 1.2 million previously unknown genes; 700 new proteorhodopsin-like photoreceptors
  - finding rhodopsin-like genes in marine microbes (requires a reassessment of oceanic carbon cycles)



Phylotypes: genomic species determined by the presence of taxonomic markers (e.g. 16S rRNA encoding genes) in the sequence



NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Search
Advanced Help

Format: Abstract Send to

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

Environmental genome shotgun sequencing of the Sargasso Sea.

Venter JC[1], 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.

Author information

Abstract

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.

Comment in

Genomics and evolution. Shotgun sequencing in the sea: a blast from the past? [Science. 2004]

PMID: 15001713 DOI: 10.1126/science.1093857

[PubMed - indexed for MEDLINE] Free full text

Publication Types, MeSH Terms, Substances

LinkOut - more resources

PubMed Commons PubMed Commons home

0 comments

How to join PubMed Commons

Full text links

Science AAAS FREE WITH REGISTRATION

Save items

Add to Favorites

Similar articles

Genomics and evolution. Shotgun sequencing in the sea: a blast from the past? [Science. 2004]

Comparative metagenomics of microbial communities. [Science. 2005]

Community genomics among stratified microbial assemblages in the ocean's interi [Science. 2006]

Review Genome analysis of marine photosynthetic microc [Curr Opin Biotechnol. 2004]

Review Status of genome projects for nonpathogenic bacteria an [Nat Biotechnol. 2000]

See reviews...

See all...

Cited by over 100 PubMed Central articles

Accelerating metagenomic read classification on CUDA-enabled GPUs. [BMC Bioinformatics. 2017]

Summer Abundance and Distribution of Proteorhodopsin Genes in [Front Microbiol. 2016]

Variations of Phosphorous Accessibility Causing Changes in Microbiome Functi [PLoS One. 2016]

See all...

Related information

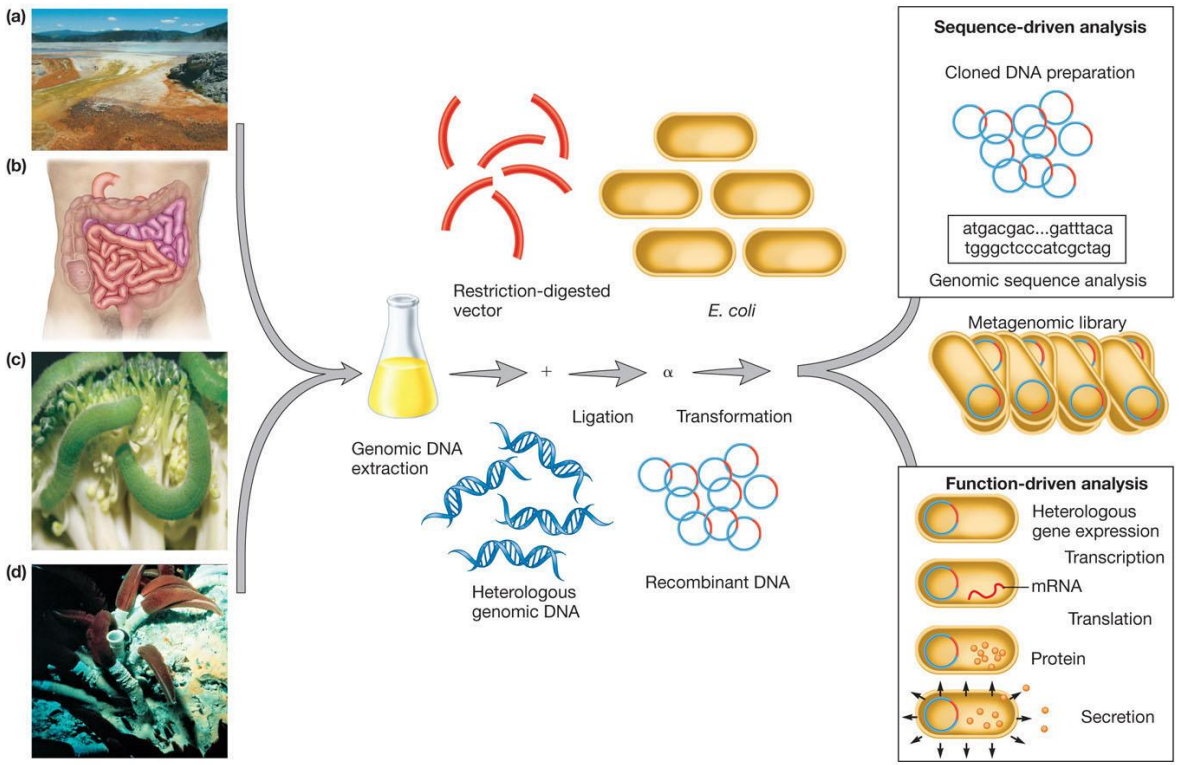
Articles frequently viewed together

Assembly

Nucleotide

Metagenomics

- used to learn more about the diversity and metabolic potential of microbial communities
  - Microbiome in humans and normal microbiota (NIH, 2007)
- **Functional metagenomics** "biotechnology"



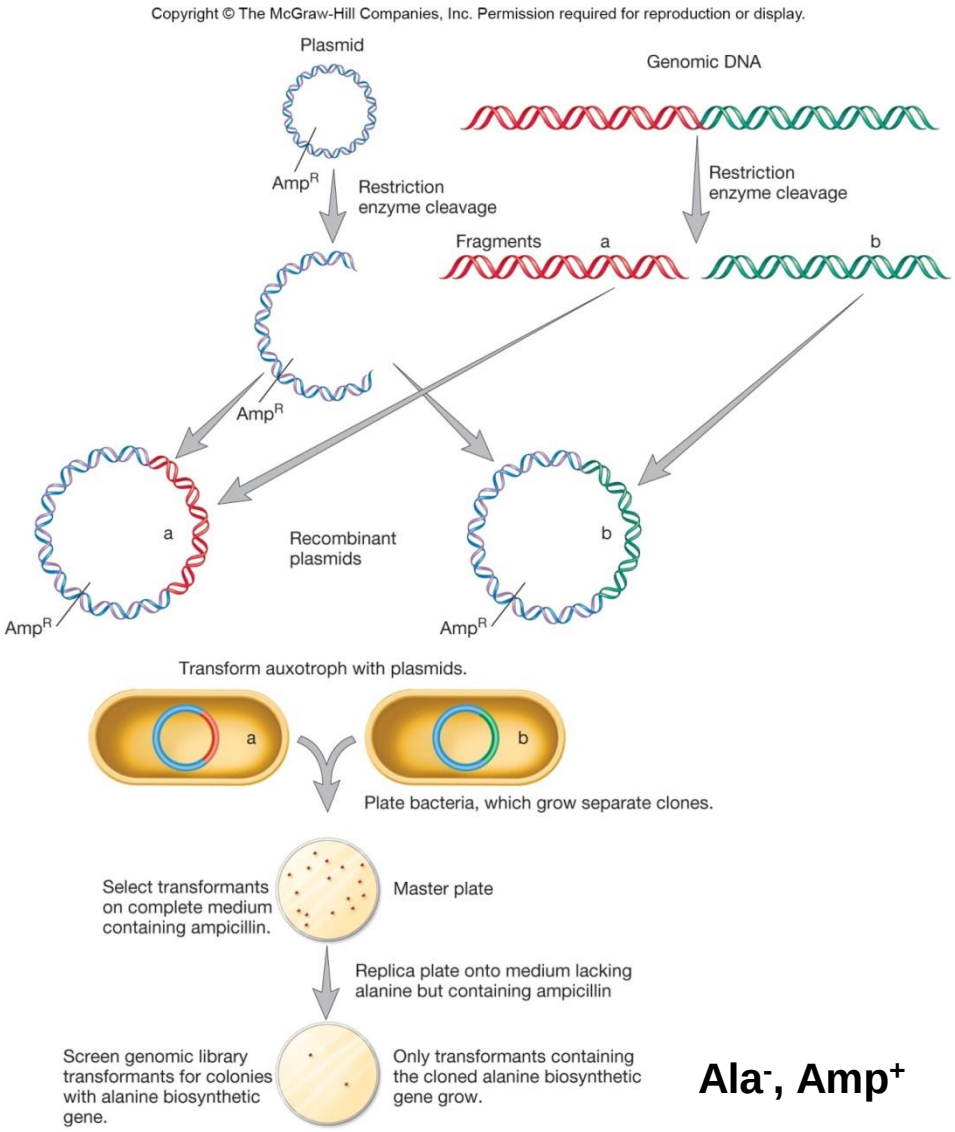
Metagenomics

Functional Metagenomics

a: Jo Handelsman, Metagenomics *MMBR*, ASM Press Dec 2004, vol., 68, fig. 3 page 674; b: © The Mc-Graw-Hill Companies, Inc.; c: Jo Handelsman, Metagenomics *MMBR*, ASM Press Dec 2004, vol., 68, fig. 3 page 674; d: © OAR/National Undersea Research Program (NURP)/College of William & Mary/NOAA

Construction of Genomic Libraries and Screening for Function

- Used when **gene of interest** is on a **chromosome that has not been sequenced**
- The library is constructed by cleaving the genome and then cloning the fragments into vectors
- The libraries are screened for the genes of interest in a variety of ways
- For example *via* **phenotypic rescue** or **genetic complementation** (e.g. screening for enzymes needed for Ala synthesis by the use of alanine auxotrophs, see Fig.)



- **III Measuring Microbial Activities in Nature**

  - **Radioisotopes (e.g. Fish-MAR)**
  - **Microelectrodes**
  - **Stable isotopes (e.g. SIP, NanoSIMS)**

Radioisotopes

Microbial activity measurements.

(a) Chemical assay for lactate and H_2S during sulfate reduction.

(b-d) Use of radioisotopes.

(b) Photosynthesis measured in natural seawater ($^{14}CO_2$).

(c) Sulfate reduction in mud ($^{35}SO_4^{2-}$).

(d) Production of $^{14}CO_2$ from ^{14}C-glucose.

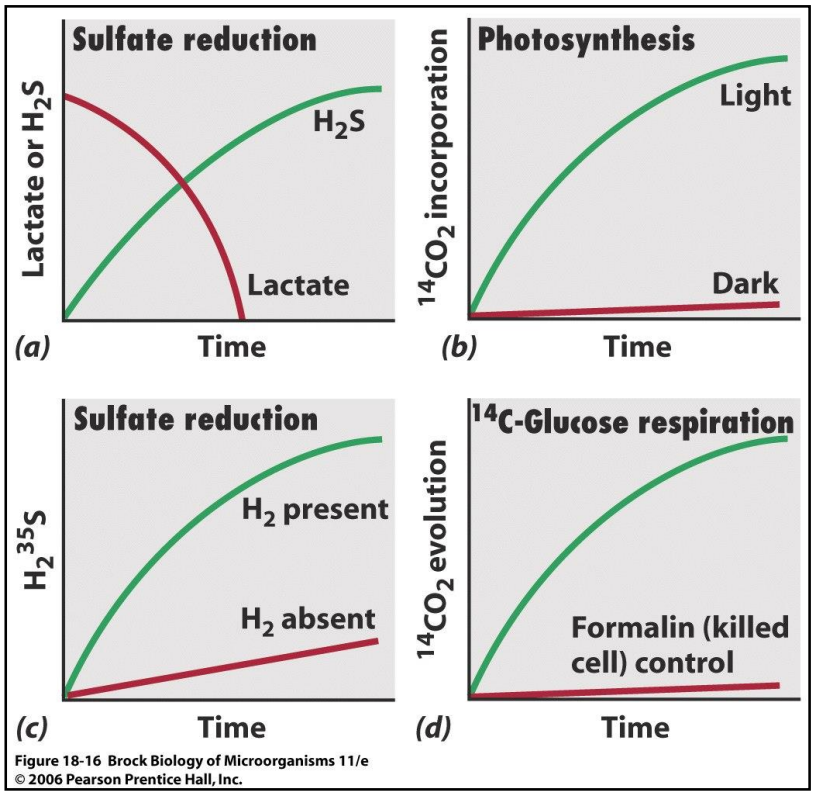


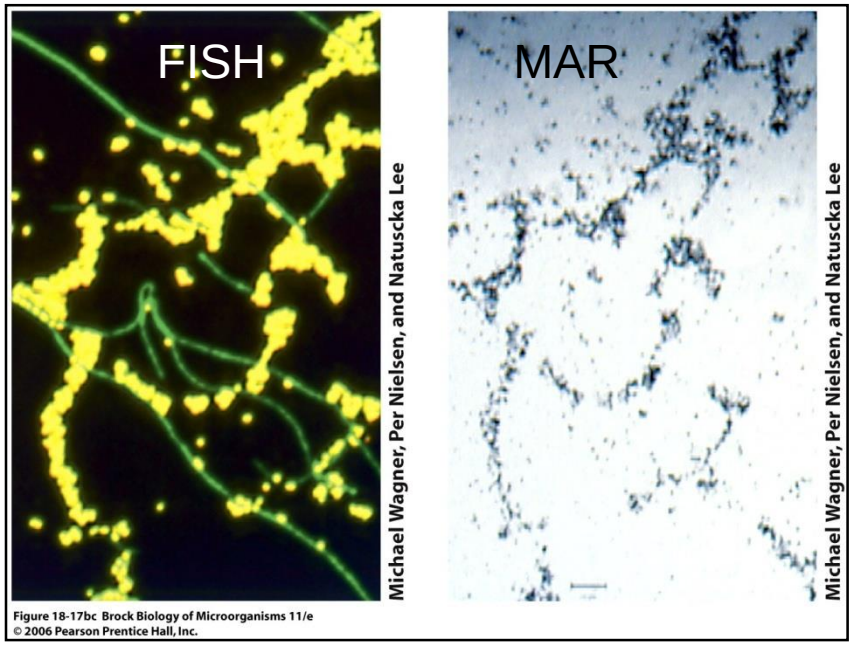
Figure 18-16 Brock Biology of Microorganisms 11/e

Fish-MAR

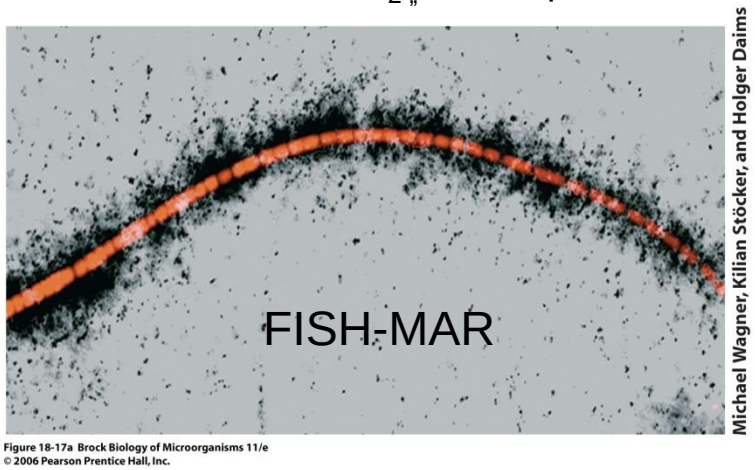
Microautoradiography (MAR)

-cells are exposed to radioactive isotope (pure culture/microbial community)
-cells affixed to slide dipped in photographic emulsion
-darkness, radioaktive decay of incorporated substrate exposes silver grains
-appear as black dots within and around the cells.
-MAR can be combined with FISH (FISH-MAR)

Mixture: ***E.coli* & *Herpetosiphon aurantiacus***
^{14}C-glucose only metabolized by *E. coli*



Uncultured cell (Gamma proteobacteria)
$^{14}CO_2$ „Autotroph“

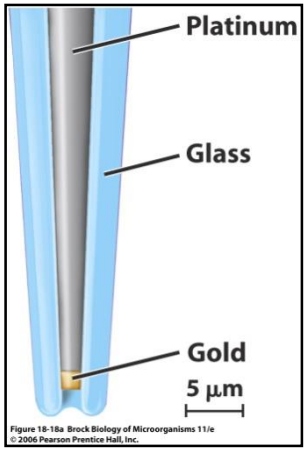


Microelectrodes

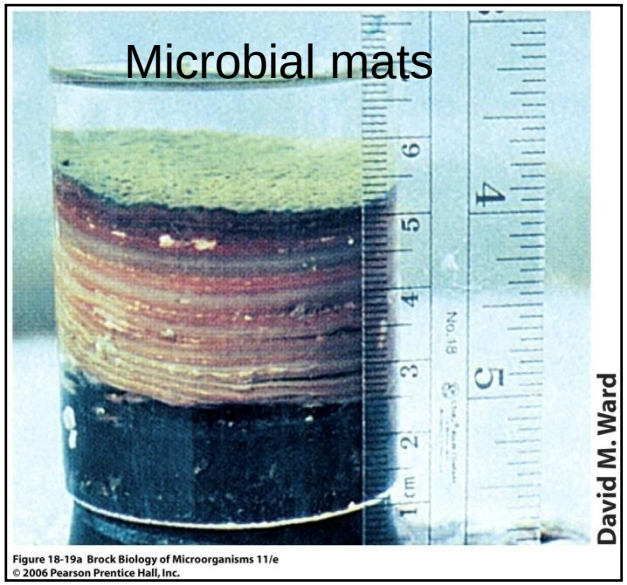
Niels Peter Revsbech

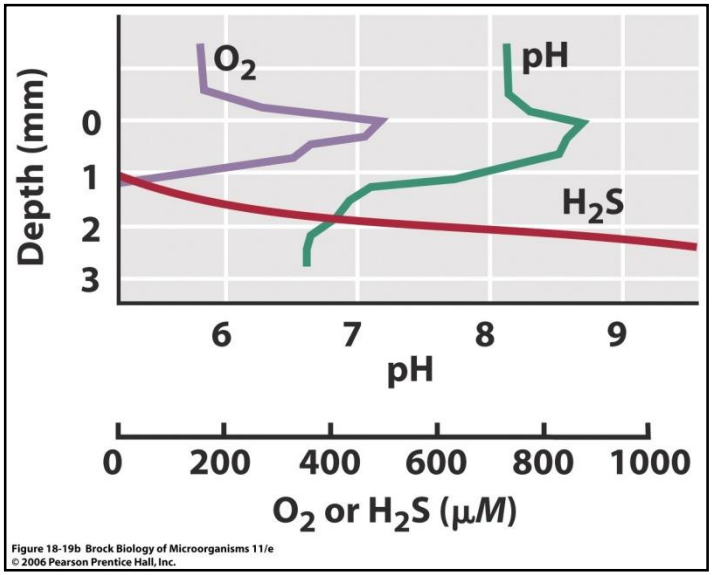
Figure 22.21

Microelectrodes



- Microelectrodes (small glas electrodes, quite fragile) to study the activity of microorganisms in nature, e.g. for measuring pH, oxygen, N_2O, CO_2, H_2 and H_2S
- Tips 2-100 µm
- Carefully inserted into the habitat (e.g. microbial mats)
- Measurements taken every 50–100µm





Layered microbial communities:

-Cyanobacteria

-anoxic phototrophic bacteria (until light-limited) & chemoorganotrophs (sulfate-reducing bacteria)

Stable Isotopes

- **For many elements several isotopes exist, differing in number of neutrons present. "Stable isotopes", not radioactive.**
  - Carbon (CO_2), Sulfur (SO_4^{2-}, H_2S)
- **Isotope fractionation** can reveal the **biological origin** of various substances.
- Fractionation is a result of the activity of enzymes that discriminate against the heavier form of an element when binding their substrates.

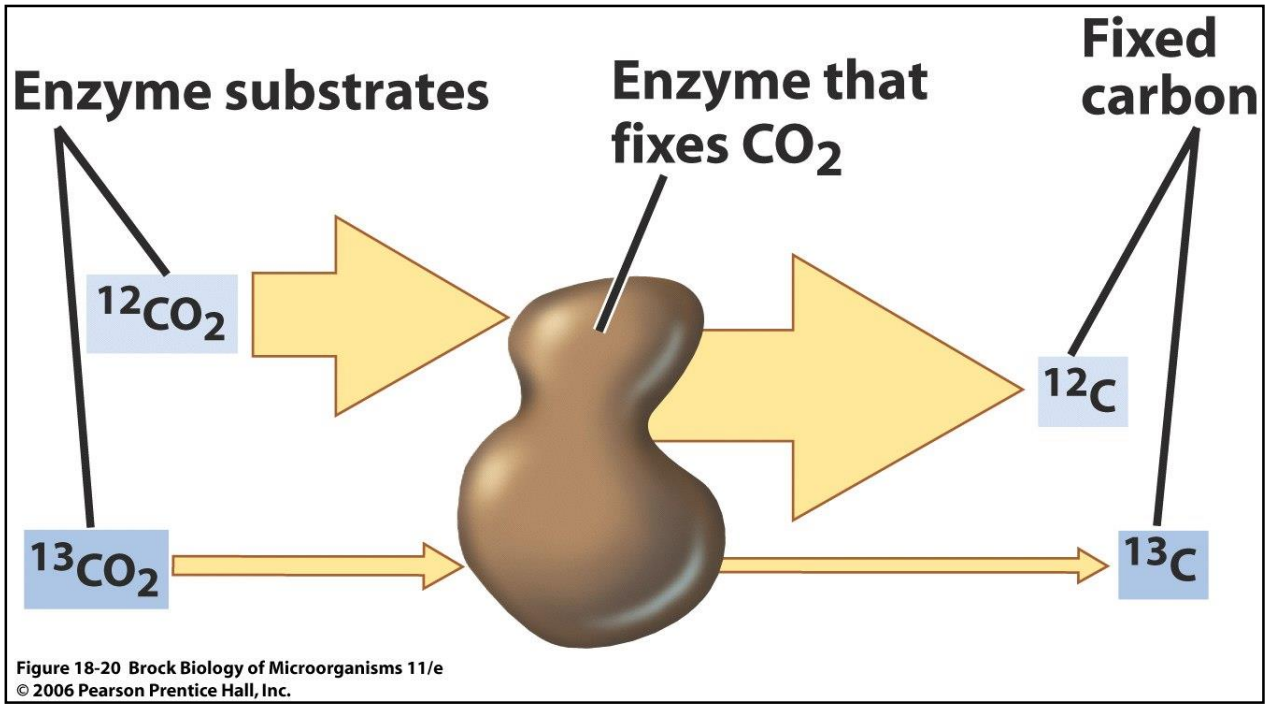


Figure 18-20 Brock Biology of Microorganisms 11/e

Stable Isotopes

Table 29.3 Natural Abundances of Some Biologically Relevant Isotopes

Hydrogen	Carbon	Nitrogen	Oxygen	Sulfur
^{1}H 99.984%	^{12}C 98.89%	^{14}N 99.64%	^{16}O 99.763%	^{32}S 95.02%
^{2}D 0.0156% (deuterium)	^{13}C 1.11%	^{15}N 0.36%	^{17}O 0.0375%	^{33}S 0.75%
			^{18}O 0.1995%	^{34}S 4.21%
				^{36}S 0.02%

Stable Isotopes

- Evidence for microbial life in ancient rocks rests on the **fossilized remains of cells** and the **isotopicaly „light“ carbon** abundant in these rocks
- **„Stromatolites“** are fossilized microbial mats

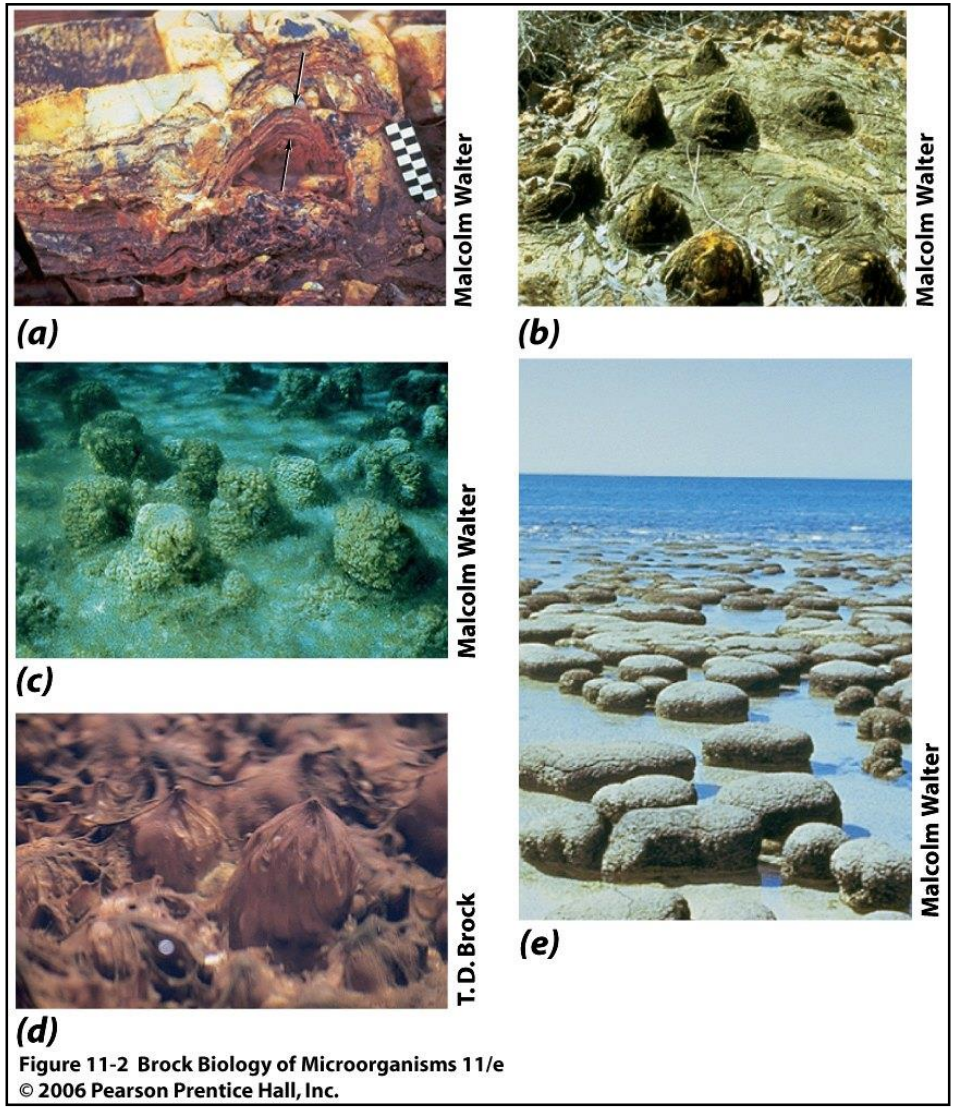


Figure 11-2 Brock Biology of Microorganisms 11/e

Stable Isotope Probing (SIP)

- Links **specific metabolic activity** to diversity using a **stable isotope**
  - Microorganisms metabolizing stable isotope (e.g., ^{13}C) **incorporate it into their DNA**
    - DNA with ^{13}C can then be used to identify the organisms that metabolized the ^{13}C
  - SIP of RNA can be done instead of DNA

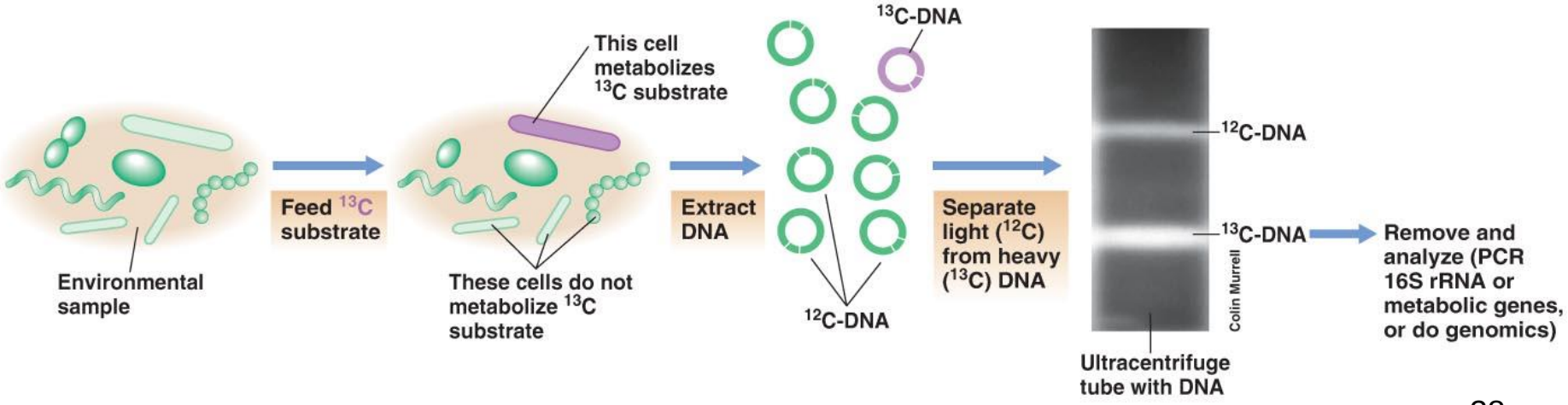
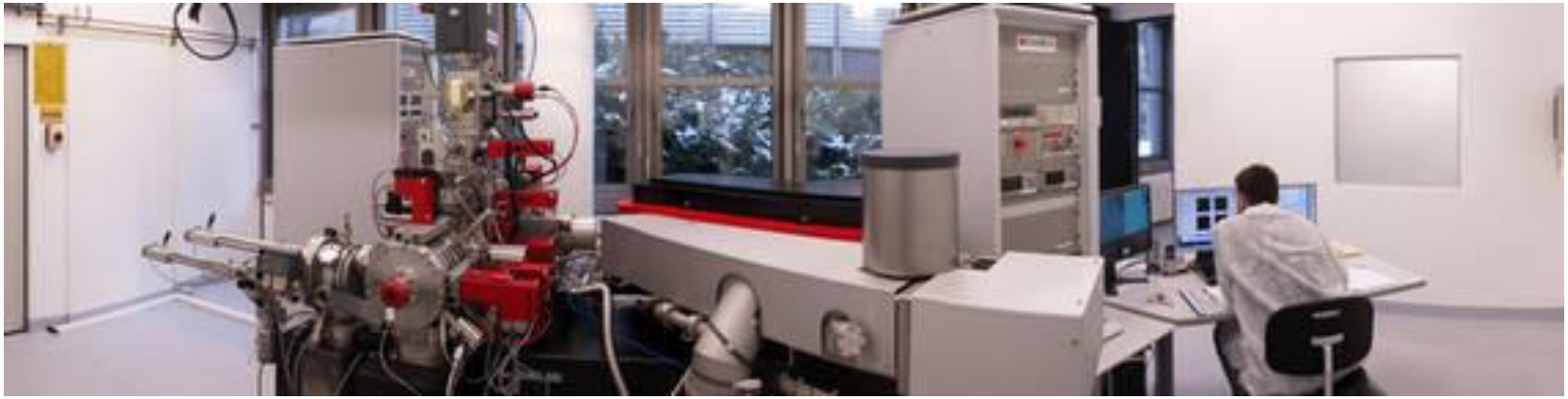


Figure 22.24

NanoSIMS

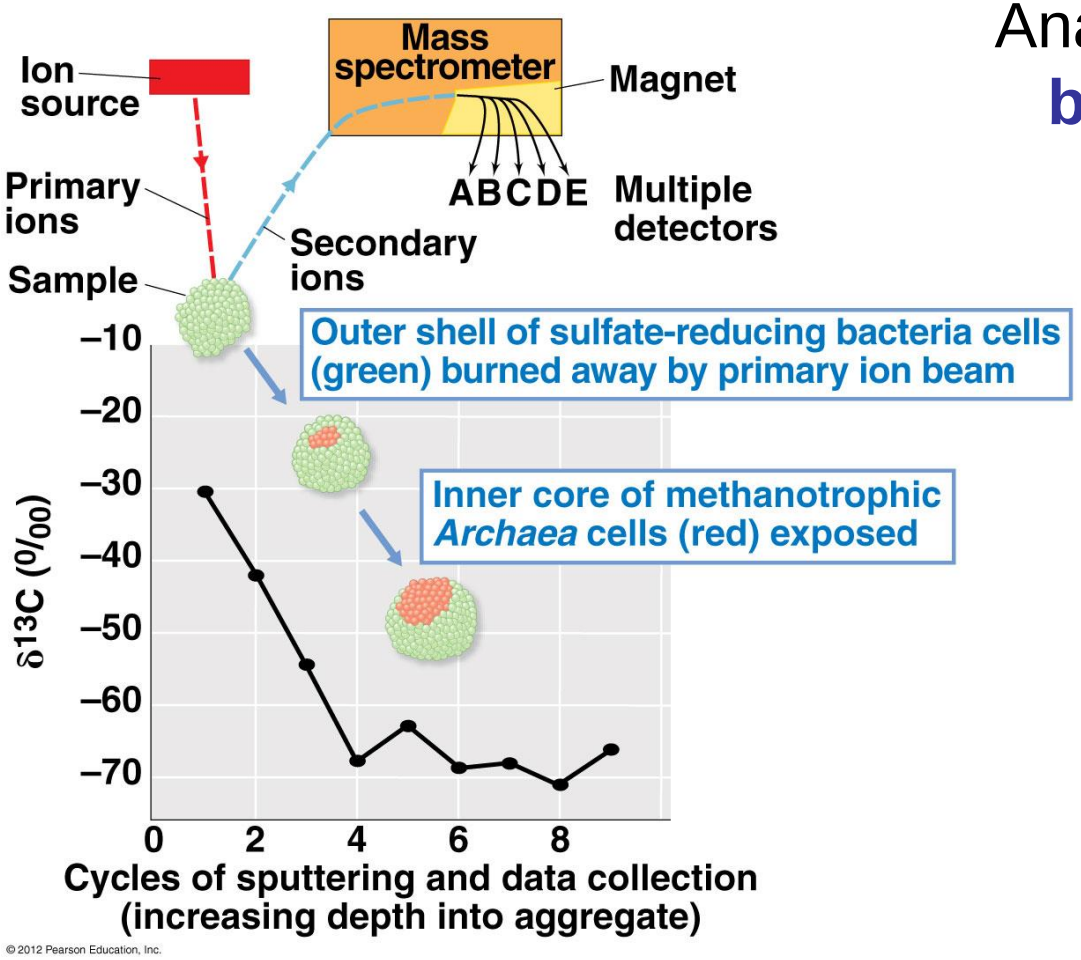


University of Vienna
Photograph by Gregor Eder
http://www.soil-science.com/index.php?id=nanosims

Secondary ion mass spectrometry (SIMS)

- Focused primary ion beam (eg. Cesium (Cs^+))
- Data on the elemental and isotopic composition of a sample

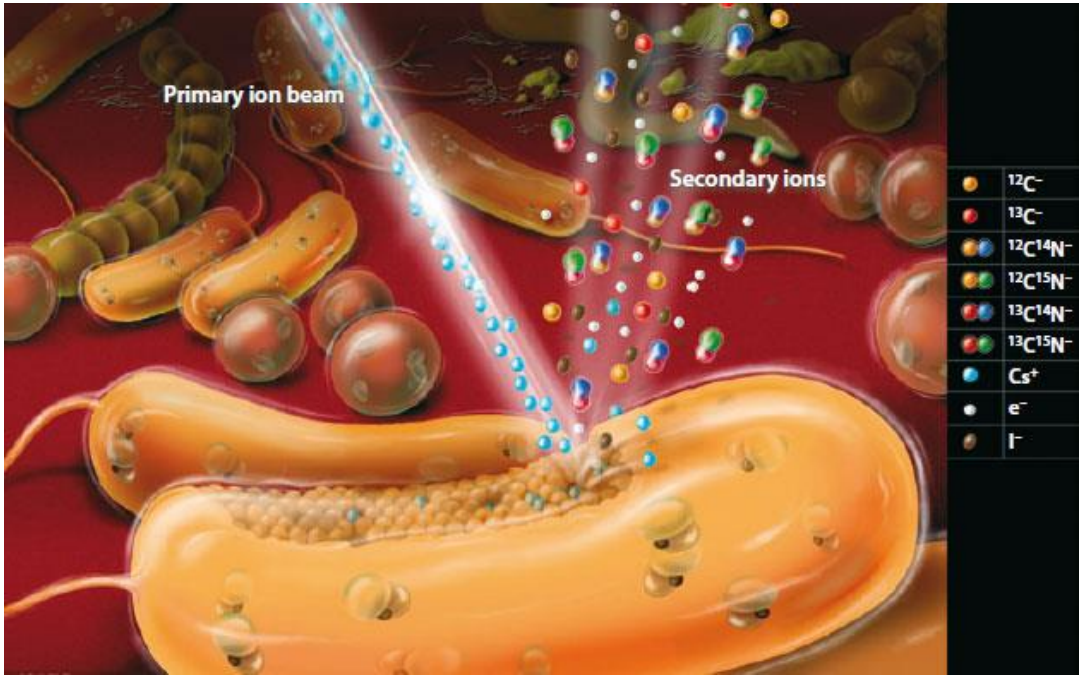
Secondary Ion Mass Spectrophotometry (SIMS)



Analysis of **single cells by SIMS**

- High-energy ion beam impacts a sample
- Secondary ions released (*sputtering*)
- Mass spectrometry of secondary ions

NanoSIMS



- Dynamic **secondary ion mass spectrometry (SIMS)** analysis of a bacterial community (not to scale).
- The bacterial cell sputtered with the primary ion beam is **labeled with ^{15}N and ^{13}C**.
- Only selected **secondary ions** are displayed. Note that **nitrogen is detected as cyanide ion** and that isobaric ions are formed (e.g., $^{12}C^{15}N^-$ and $^{13}C^{14}N^-$), whose differentiation requires high mass resolution as offered, for example, by modern NanoSIMS instruments.
- In addition, the analyzed rod-shaped cell is also labeled with an **iodized oligonucleotide probe targeting the ribosomal RNA**. If sufficient cellular material is removed by sputtering to allow secondary ion formation from the ribosomes (displayed in the cell lumen), iodine ions (I^-) can also be detected in the secondary ions.
- Figure by Gerhard Pucher, http://www.soil-science.com/index.php?id=nanosims

A single-cell view on the ecophysiology of anaerobic phototrophic bacteria

Niculina Musat[a,1], Hannah Halm[a], Bärbel Winterholler[b], Peter Hoppe[b], Sandro Peduzzi[c], Francois Hillion[d], Francois Horreard[d], Rudolf Amann[a], Bo B. Jørgensen[a], and Marcel M. M. Kuypers[a]

[a]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany; [b]Max Planck Institute for Chemistry, Johann-Joachim-Becher Weg 27, 55128 Mainz, Germany; [c]Cantonal Institute of Microbiology and Alpine Biology Center Foundation Piora, Via Mirasole 22A, CH-6500 Bellinzona, Switzerland; and [d]Cameca, Quai des Gresillons 29, 92622 Gennevilliers Cedex, France

Quantitative information on the ecophysiology of individual microorganisms is generally limited because it is difficult to assign specific metabolic activities to identified single cells. Here, we develop and apply a method, Halogen In Situ Hybridization-Secondary Ion Mass Spectroscopy (HISH-SIMS), and show that it allows simultaneous phylogenetic identification and quantitation of metabolic activities of single microbial cells in the environment. Using HISH-SIMS, individual cells of the anaerobic, phototropic bacteria *Chromatium okenii*, *Lamprocystis purpurea*, and *Chlorobium clathratiforme* inhabiting the oligotrophic, meromictic Lake Cadagno were analyzed with respect to $H^{13}CO_3^-$ and $^{15}NH_4^+$ assimilation. Metabolic rates were found to vary greatly between individual cells of the same species, showing that microbial populations in the environment are heterogeneous, being comprised of physiologically distinct individuals. Furthermore, *C. okenii*, the least abundant species representing ≈0.3% of the total cell number, contributed more than 40% of the total uptake of ammonium and 70% of the total uptake of carbon in the system, thereby emphasizing that numerically inconspicuous microbes can play a significant role in the nitrogen and carbon cycles in the environment. By introducing this quantification method for the ecophysiological roles of individual cells, our study opens a variety of possibilities of research in environmental microbiology, especially by increasing the ability to examine the ecophysiological roles of individual cells, including those of less abundant and less active microbes, and by the capacity to track not only nitrogen and carbon but also phosphorus, sulfur, and other biological element flows within microbial communities.

Lake Cadagno (Italian: *Lago di Cadagno*) is a **oligotrophic, meromictic** lake in the Piora valley (canton of Ticino (Tessin), Switzerland.

^{15}N-ammonium and ^{13}C-inorganic carbon uptake by individual *C. okenii* cells (Fig. 5)

Musat N et al. PNAS 2008;105:17861-17866

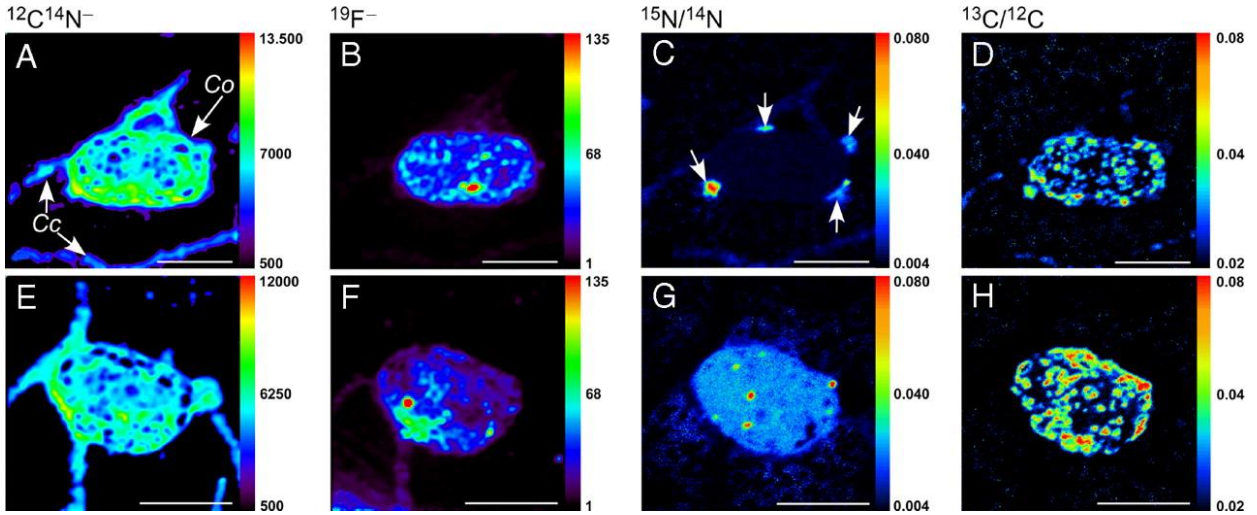


Fig. 5 Parallel secondary ion images of *C. okenii* cells (Co) and presumably *C. clathratiforme* cells (Cc) (A–H). The abundances of $^{12}C^{14}N^-$ (A and E), $^{19}F^-$ (B and F), and the $^{15}N/^{14}N$ (C and G) and $^{13}C/^{12}C$ (D and H) ratios are shown. Arrows in C indicate highly active epibionts attached to a *C. okenii* cell.

- **Halogen In Situ Hybridization-Secondary Ion Mass Spectroscopy (HISH-SIMS).** It uses horseradish-peroxidase-labeled oligonucleotide probes and fluorine-containing tyramides for the identification of micro-organisms in combination with stable-isotope-labeling experiments for analyzing the metabolic function of single microbial cells.
- **Simultaneous phylogenetic identification and quantitation of metabolic activities** of single microbial cells in the environment.
- Individual cells of the anaerobic, phototropic bacteria *Chromatium okenii*, *Lamprocystis purpurea*, and *Chlorobium clathratiforme* inhabiting the **oligotrophic, meromictic Lake Cadagno** were analyzed with respect to **$H^{13}CO_3^-$ and $^{15}NH_4^+$ assimilation**.
- For this purpose, water samples from 11.5-m depth were incubated for 12 h at *in situ* light and temperature conditions after addition of **^{15}N-labeled ammonium** and **^{13}C-labeled bicarbonate**
- For each individual cell, we recorded simultaneously secondary-ion images of naturally abundant ^{12}C (measured as $^{12}C^-$) and ^{14}N (measured as $^{12}C^{14}N^-$) atoms and, similarly, of ^{13}C, ^{15}N, and ^{19}F
- **$^{15}N/^{14}N$ and $^{13}C/^{12}C$ ratios** provide evidence for **metabolic activity**
- **^{19}F phylogenetic identification**

Ammonium and inorganic carbon assimilation by green and purple sulfur bacteria in Lake Cadgno (Fig. 6)

Musat N et al. PNAS 2008;105:17861-17

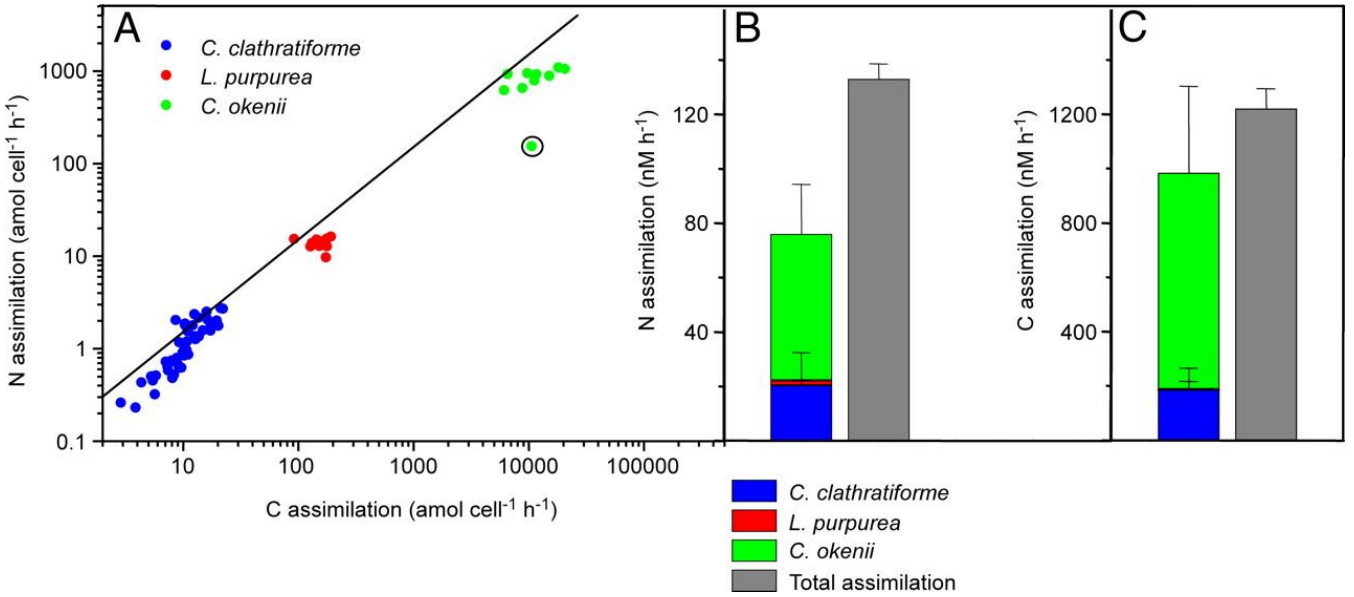


Fig. 6 Ammonium versus inorganic carbon uptake by individual *C. clathratiforme, L. purpurea*, and *C. okenii* cells (A). The line represents the theoretical 'Redfield', C:N ratio of 6.6. Contribution to the total ammonium (B), and total dissolved inorganic carbon (C) assimilation in the system by each population. For the calculation of total nitrogen and carbon uptake, the absolute abundance of each species as determined by CARD-FISH was taken into account.

- Metabolic rates were found to vary greatly between individual cells of the same species (**heterogeneous population of physiologically distinct individuals**).
- Furthermore, ***C. okenii*, the least abundant species representing 0.3% of the total cell number**, contributed **more than 40% of the total uptake of ammonium and 70% of the total uptake of carbon in the system**
- **Less abundant microbes can play a significant role in the nitrogen and carbon cycles in the environment**

? Questions ?

- Which three PCR based methods used in microbial ecology do you know (clone library, denaturing gradient gel electrophoresis „DGGE“, Terminal Restriction Fragment Length Polymorphism “T-RFLP”)? Describe the methods.
- How does a phylochip analysis work?
- Explain the term „metagenomics“. What can be learned from this technology?
- Describe a typical metagenomic approach. How does it work?
- Describe the construction of a metagenomic clone libarary
- Describe two methods how microbial activities can be determined in Nature (e.g. Fish-MAR, microelectrodes, SIP).

Genomics

- **The discipline involving mapping, sequencing, analyzing, and comparing genomes.**
- **The study of all of the nucleotide sequences, including structural genes, regulatory sequences, and noncoding DNA segments, in the chromosomes of an organism.**

Bacteria, *Escherichia coli*

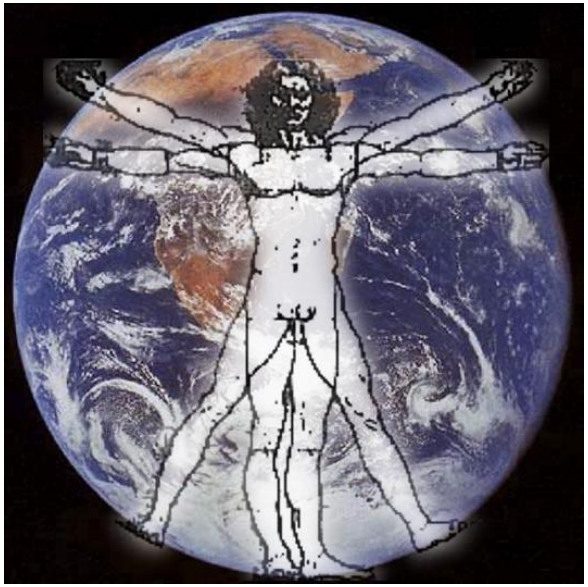
Eucarya

What is genomics ?

- A marriage of molecular biology, robotics, and computing
- Tools and techniques of recombinant DNA technology
  - e.g., DNA sequencing
- High-throughput technology
  - e.g., robotics for sequencing
- Computers are essential for processing and analyzing the large quantities of data generated

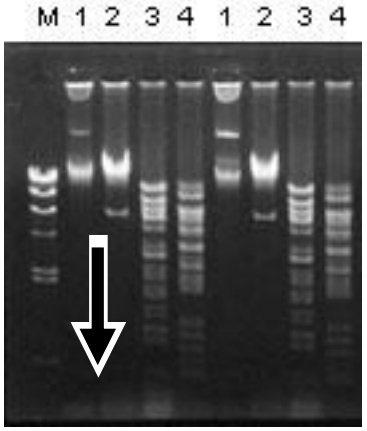
Origins of genomics

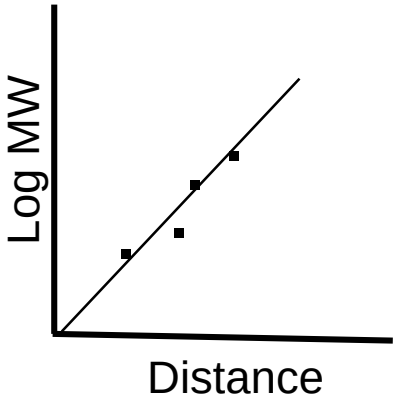
- Human Genome Project
  - Goal: sequence 3 billion (Mrd.) base pairs
  - High-quality sequence
- Immensity of task required new technologies
  - Automated sequencing
- Decision to sequence other genomes
  - Beginnings of comparative genomics

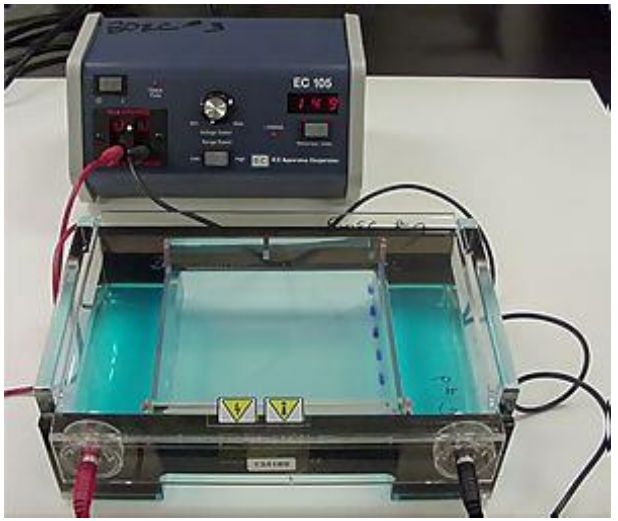


Technical foundations of genomics

- Molecular biology: recombinant-DNA technology
- DNA sequencing
- Library construction
- PCR amplification
- Hybridization techniques







Steps in genomic sequencing

I Library making

- Small/Large-insert library from genome

II Production sequencing

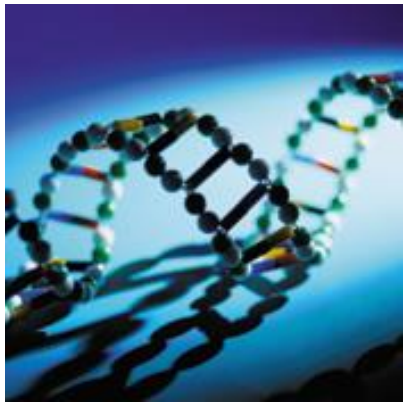
- Generate fragments to be sequenced
- Perform sequencing reactions
- Determine sequence

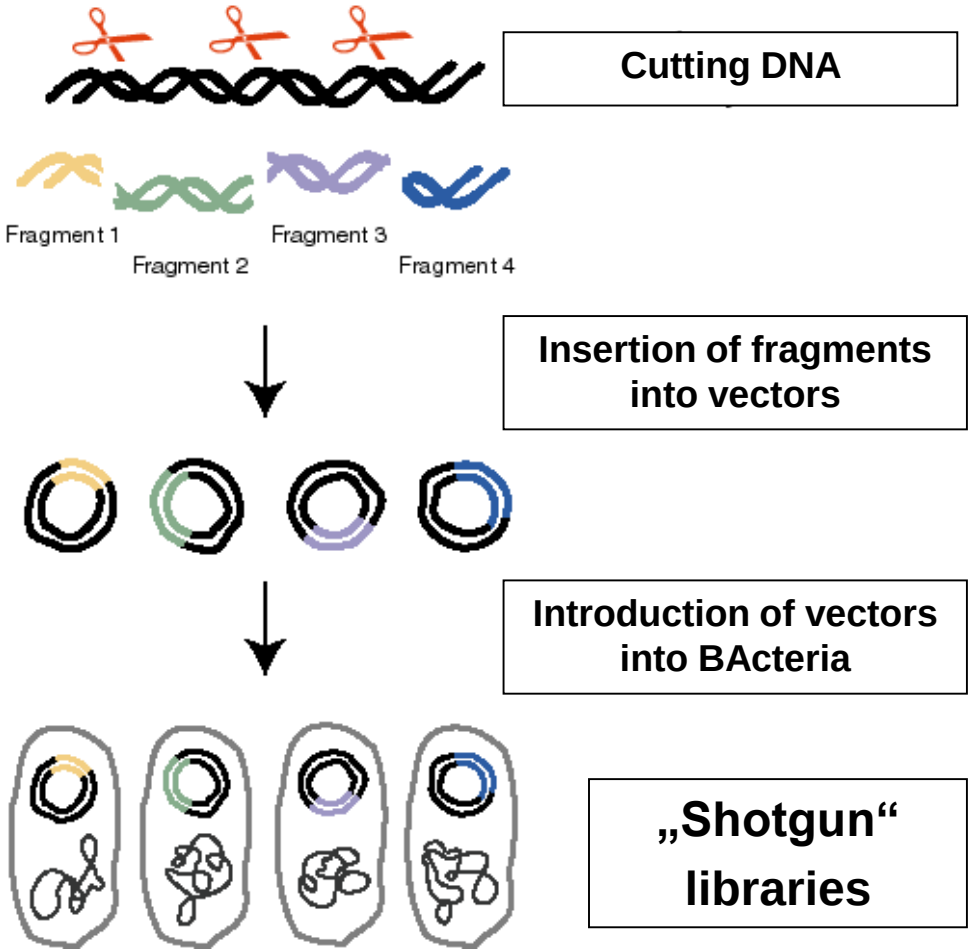
III Finishing

- Assemble into continuous sequence
- Fill gaps

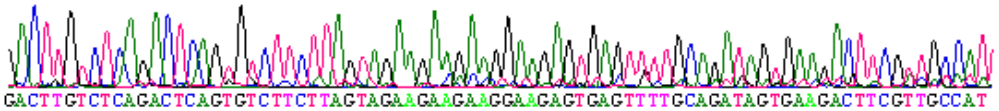
IV Annotation

„Shotgun Sequencing“





„Shotgun Sequencing“



Shotgun sequencing of libraries

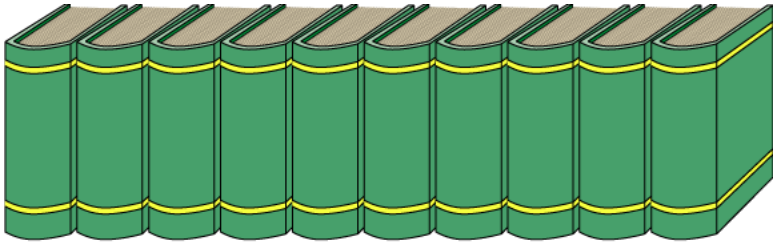
Assembley

Contig Contig

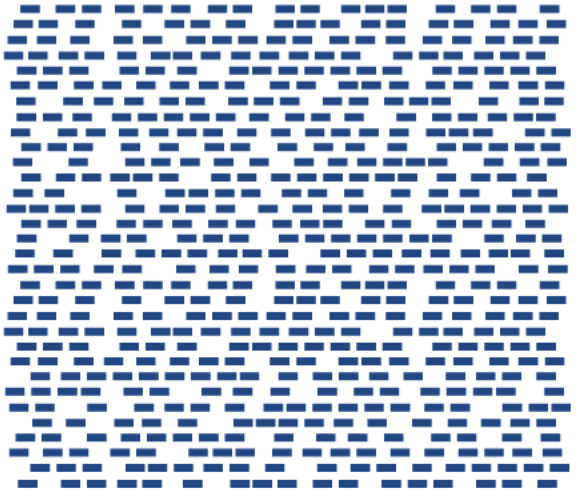
Gap closure

Genome

Whole-genome shotgun sequencing



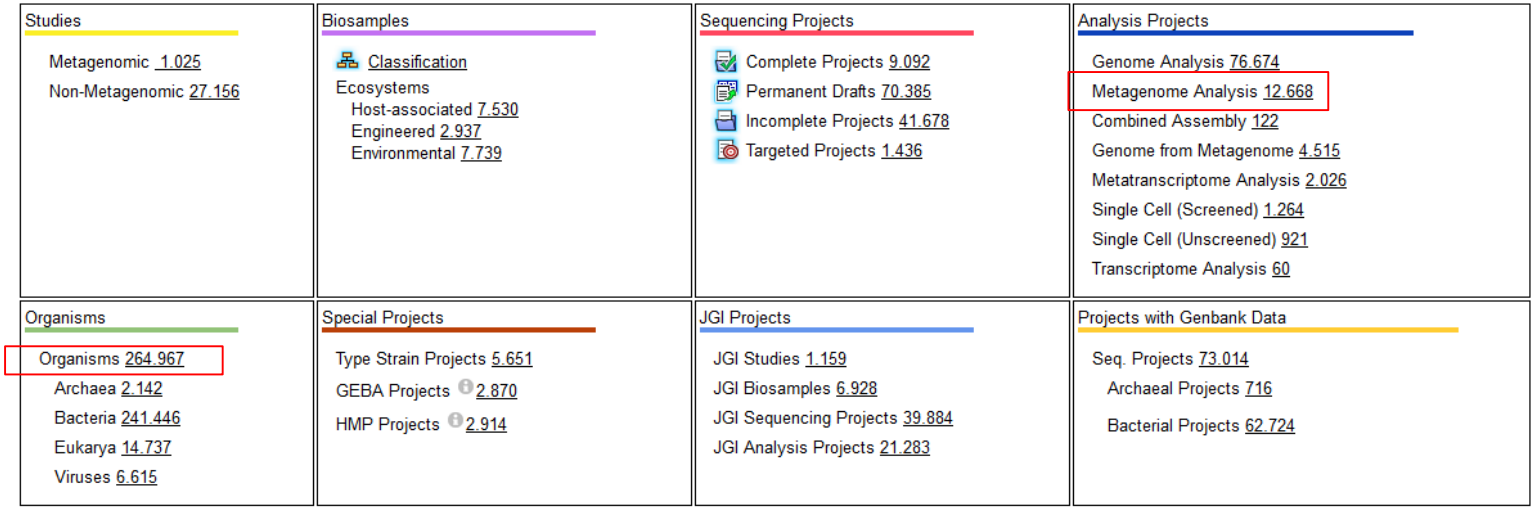
Generate tens of millions of sequence reads



Assemble

DNA sequencing

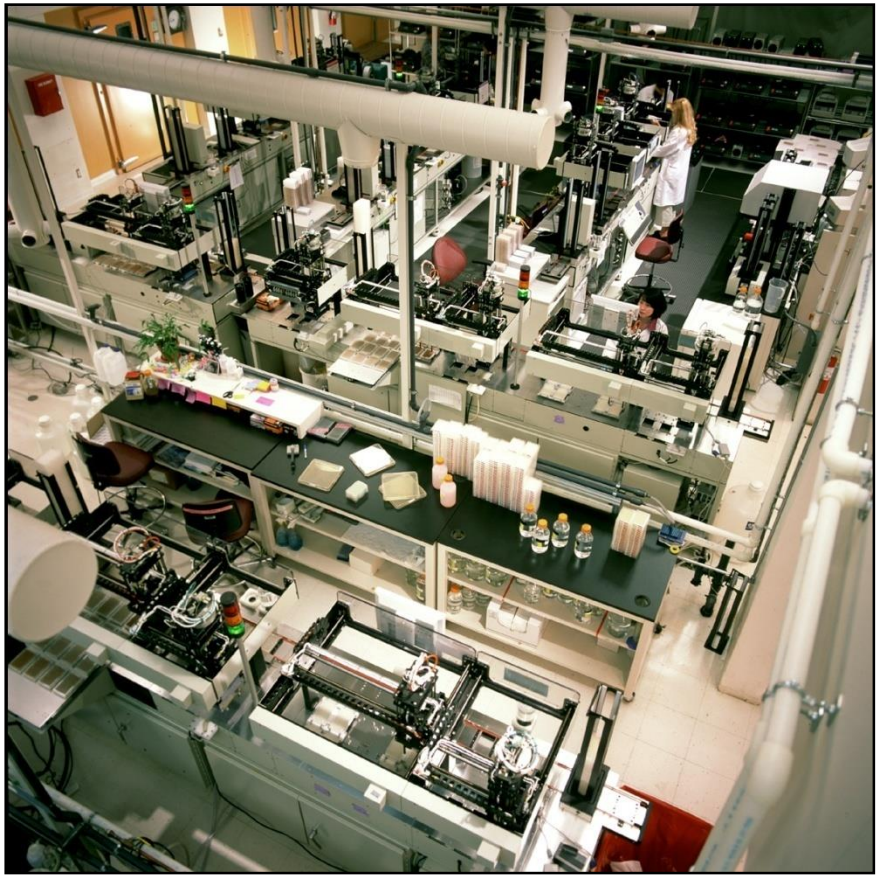
- Currently, there are **264.967 genomes sequenced and published** (**many ongoing genome projects**).
  - Provides lots of information about potential of organism
- In metagenomics the DNA from an environment is sequenced (**12.668 metagenome projects**).
  - Sargasso Sea sequencing project by C. Venter



Please cite:
Supratim Mukherjee; Dimitri Stamatis; Jon Bertsch; Galina Ovchinnikova; Olena Verezemska; Michelle Isbandi; Alex D. Thomas; Rida Ali; Kaushal Sharma; Nikos C. Kyrpides; T. B. K. Reddy. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements *Nucl. Acids Res. (2016); doi: 10.1093/nar/gkw992*
Full text

(ii) Industrialization of sequencing

- Most large-scale sequencing projects divide tasks among different teams
  - Large-insert libraries
  - Production sequencing
  - Finishing
- Sequencing machines run 24/7
- Many tasks performed by robots



Genomics, P. N. Benfey & A. D. Protopapas, Pearson/Prentice Hall

III Finishing

Sequence assembly

- Random sequences
  - First assemble into overlapping sequence
  - Then create one continuous sequence
- Program used for this operation named PHRAP
  - Analyzes each position to determine the following:
    - Quality of sequence
    - Consistency of sequence of same region
      - Acquired from different random fragments

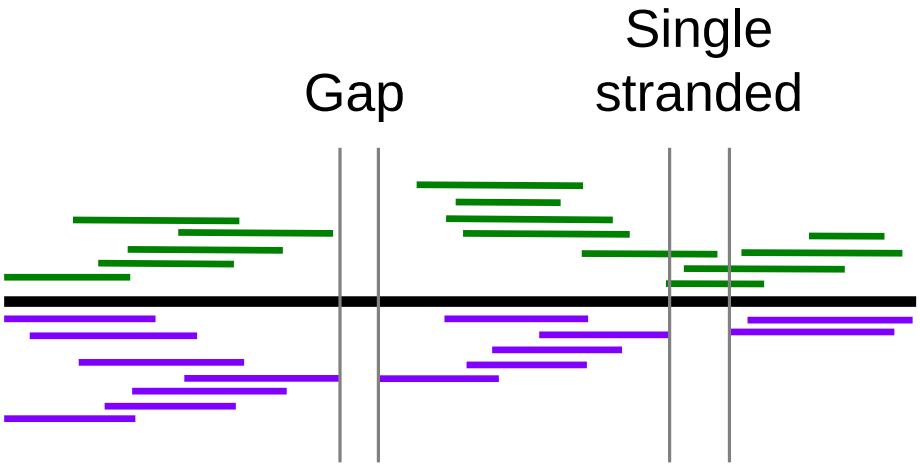
Sequence assembly readout

- The readout generated by the PHRAP program shows overlapping sequences lined up to form one contiguous sequence.

Genomics, P. N. Benfey & A. D. Protopapas, Pearson/Prentice Hall

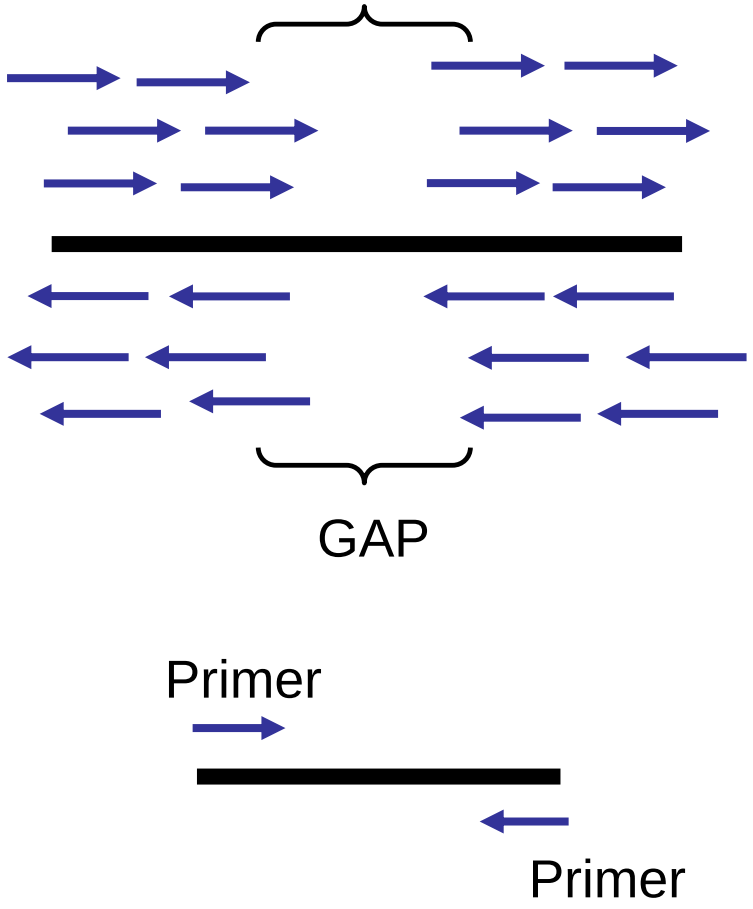
Finishing I

- Process of assembling raw sequence reads into accurate contiguous sequence
  - Required to achieve 1/10,000 accuracy
- Manual process
  - Look at sequence reads at positions where programs can't tell which base is the correct one
  - Fill gaps
  - Ensure adequate coverage



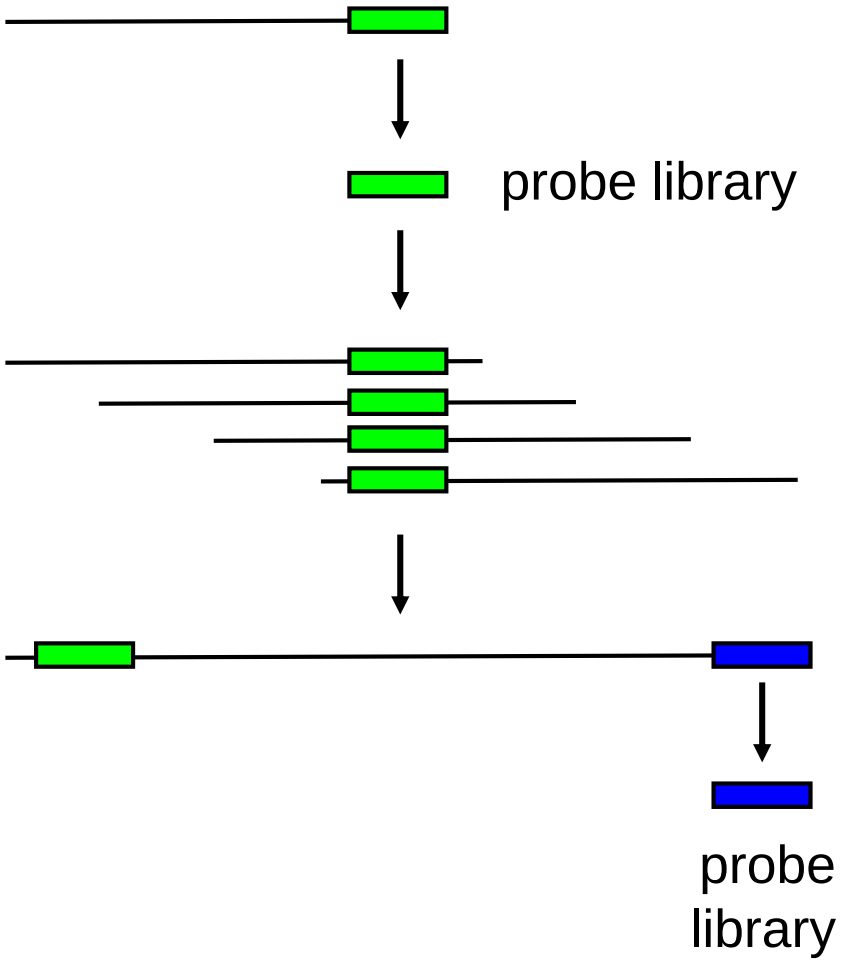
Finishing II

- To fill gaps in sequence, design primers and sequence from primer
- To ensure adequate coverage, find regions where there is not sufficient coverage and use specific primers for those areas



Chromosome walking

- Combines probing with insert ends and restriction mapping
- First find hybridizing clones
  - Then create a restriction map
- Identify the clone with the shortest overlap
- Make probe from its end
- Repeat process



Genomics, P. N. Benfey & A. D. Protopapas, Pearson/Prentice Hall

Verification

- Region verified for the following:
  - **Coverage** (i.e., how many times the same region has been sequenced, and in what direction)
  - **Sequence** quality (i.e., whether ambiguity has been removed for all positions in the sequence)
  - **Contiguity** (i.e., whether the sequence forms one uninterrupted stretch of DNA)
- Determine restriction-enzyme cleavage sites
  - Generate **restriction map** of sequenced region
  - Must agree with **fingerprint generated of clone** during mapping step

IV Annotation

Bioinformatics

- Analysis of genome data using **computers**
- Generates data on **genome content, structure, and arrangement**
- Also provides data on **protein structure and function**
- Uses annotation to determine **location of genes** on newly sequenced genome
- Further examination carried out using ***in silico* analysis**

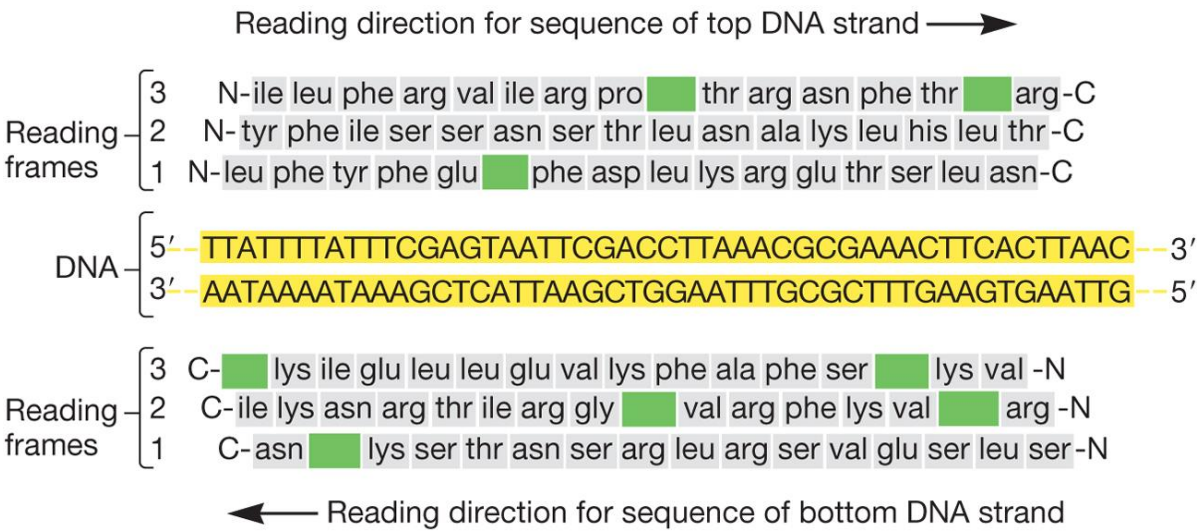
Annotating the Genome

After finishing and closing the genome sequence computers search for:

- **Open reading frames (ORFs) and genes encoding protein homologues**
- **Exons and introns**
- **Regulatory sequences**
- **Repetitive elements**

Definition ORF: a sequence of DNA that encodes no in-frame stop codons and, if transcribed, could be translated to yield a protein of known length and composition. A *functional* ORF is one that actually encodes a protein in the cell. (Brock)

Genome Annotation - 1



- Process that locates genes in the genome map
- Identifies each open reading frame in genome
  - a reading frame >100 codons that is not interrupted by a stop codon
  - a ribosomal binding site at the 5’ end and terminator sequences at the 3’ end

Annotation: Locating ORFs

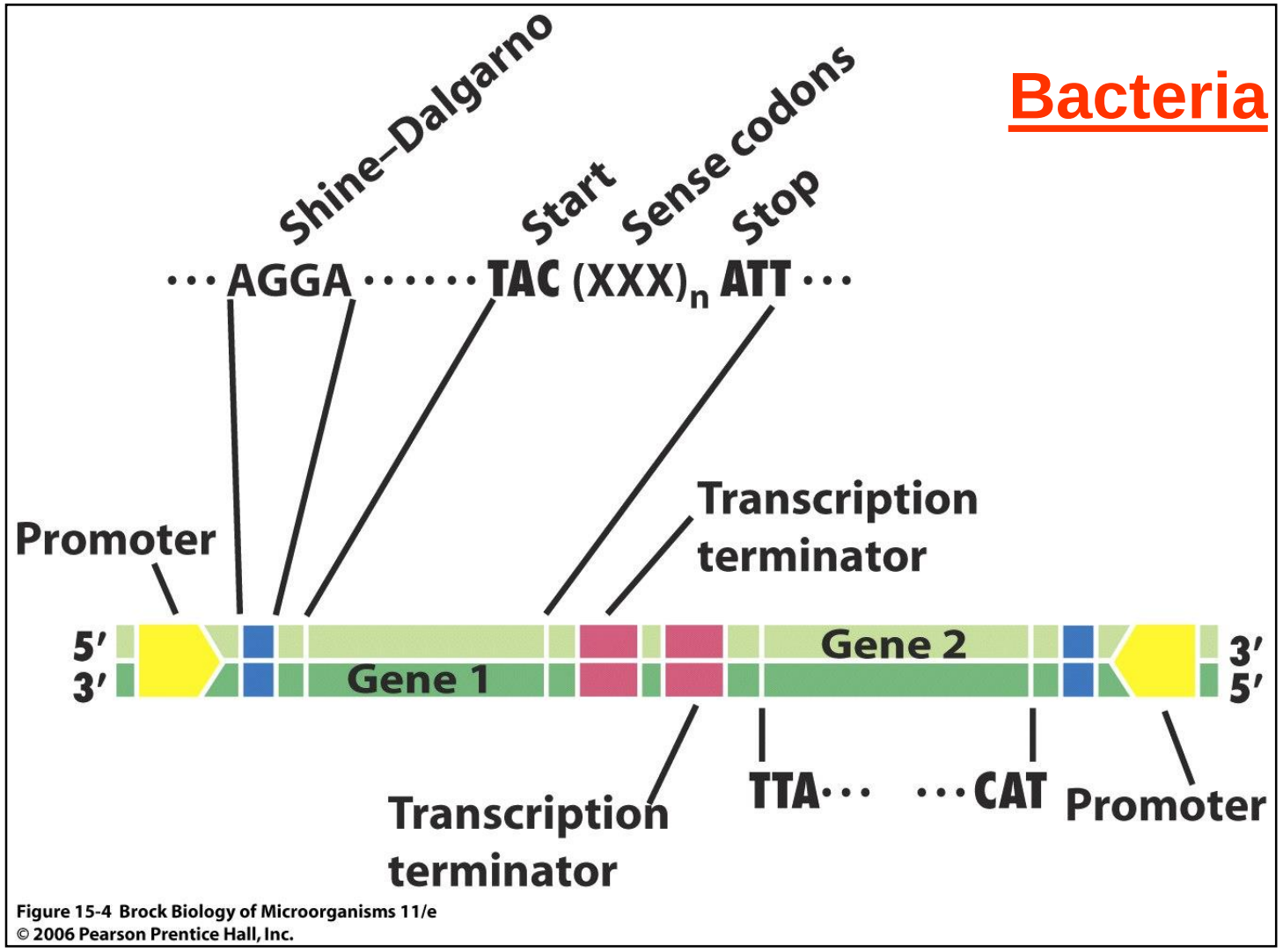
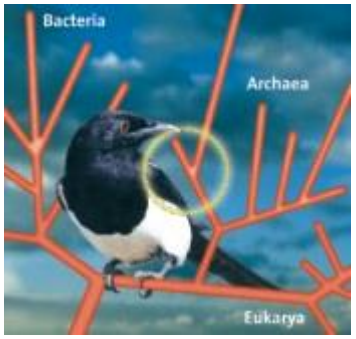


Figure 15-4 Brock Biology of Microorganisms 11/e

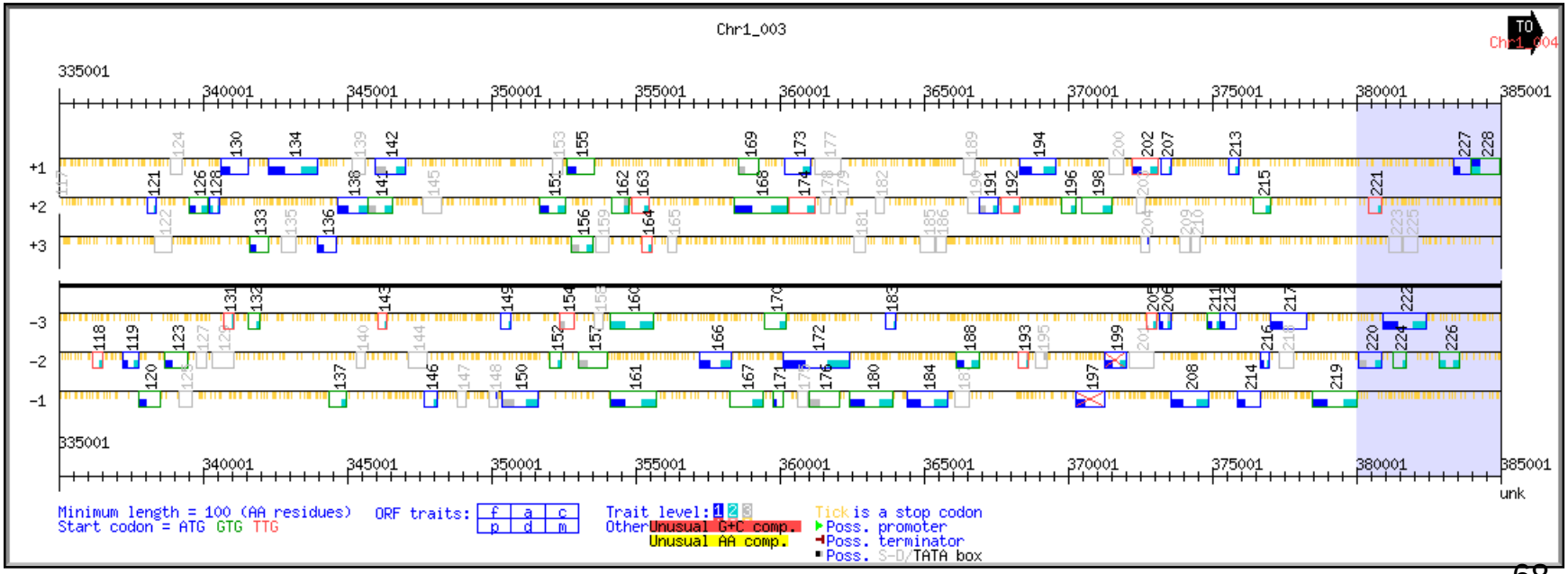
Editing and Annotation



MAGPIE Automated Genomics Project Investigation Environment

Magpie Project - Thermoproteus_tenax - **Private View**

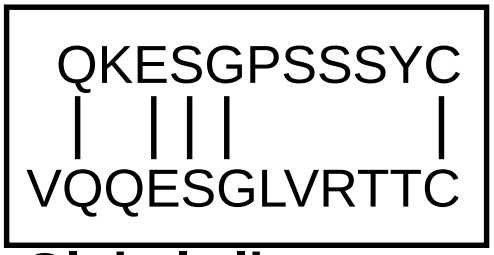
Web updated Oct 17 08:43:02 2003
Sequence last updated Mon May 12 11:25:47 2003



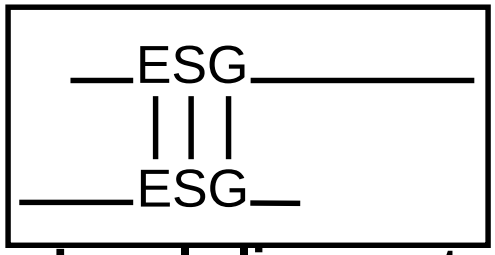
Homology searches

- **Homology** between protein or DNA sequences is defined in terms of shared ancestry (homologous proteins/sequences)
- **Search databases of DNA sequences**
- Use computer algorithms to align sequences
  - **Don't require perfect matches** between sequences
    - **Allow for insertions, deletions, and base changes**
- Most commonly used algorithms:
  - **BLAST**
  - FASTA

BLAST = Basic Local Alignment Search Tool (Altschul et al. 1990)
-National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/ BLAST/)



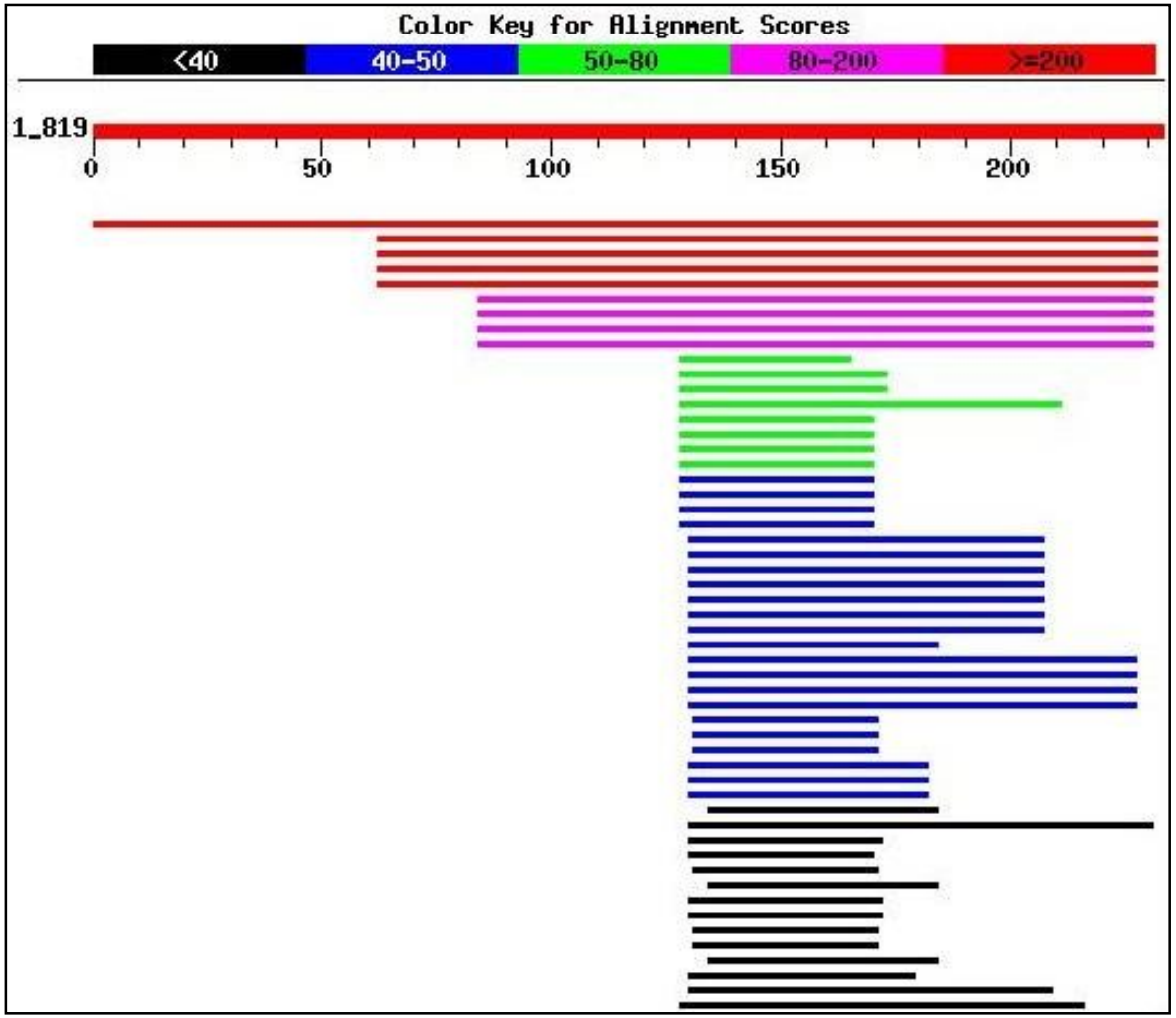
Global alignment



Local alignment

- **Local alignment tool**, attempts to find isolated regions in sequence pairs that have high level of similarity
- User provides ***query* sequence**, that is compared to the entire database (all sequences in the database = target sequences)
- **Pairwise alignment**
  (more sequences = Phylogeny)

BLAST results (hits)



- The graphical display in the slide shows the sequences that BLAST was able to align to the MASH-1 amino acid sequence.
- Alignment scores are represented on the color bar at the top of the figure, with scores going from low (black) to high (red).
- The numbered line below the color bar represents the amino acid sequence of MASH-1.
- Below it are various sequences from several databases that were found to align to MASH-1. The precise position of each sequence relative to the MASH-1 sequence indicates the areas of sequence similarity.

Detailed BLAST results

Sequences producing significant alignments:	Score (bits)	E Value
gi|112189|pir||S11563 probable MASH-2 protein - rat >gi|227...	291	8e-79
gi|440957|gb|AAB28830.1| Achaete-Scute homolog Mash-1 gene ...	283	3e-76
gi|2134688|pir||A48279 achaete scute protein - human >gi|30...	283	3e-76
gi|20455478|sp|P50553|ASH1_HUMAN Achaete-scute homolog 1 (H...	283	3e-76
gi|6678806|ref|NP_032579.1| achaete-scute complex homolog-l...	278	7e-75
gi|2642465|gb|AAB86993.1| Achaete-Scute homologue 2 [Homo s...	105	2e-22
gi|112188|pir||S11562 probable MASH-1 protein - rat >gi|566...	92	2e-18
gi|17432908|sp|O35885|ASH2_MOUSE Achaete-scute homolog 2 (M...	90	5e-18
gi|8574075|emb|CAB94773.1| Mash2 protein [Mus musculus] >gi...	89	1e-17
gi|1754729|gb|AAB39362.1| ASCL2 [Homo sapiens]	65	3e-10
gi|17456298|ref|XP_062690.1| similar to putative bHLH trans...	55	2e-07
gi|20863265|ref|XP_137216.1| similar to transcription facto...	53	1e-06
gi|27717809|ref|XP_235013.1| similar to Achaete-scute homol...	52	1e-06
gi|27679426|ref|XP_215039.1| similar to putative bHLH trans...	52	2e-06
gi|18249653|dbj|BAB83912.1| putative bHLH transcription fac...	51	3e-06
gi|28273166|tpg|DAA00301.1| TPA: class II basic helix-loop-...	51	3e-06
gi|20910395|ref|XP_136181.1| similar to putative bHLH trans...	50	4e-06
gi|13928056|emb|CAC37689.1| MASH5 protein [Mus musculus] >g...	50	7e-06
gi|18249655|dbj|BAB83913.1| putative bHLH transcription fac...	49	2e-05
gi|10190680|ref|NP_065697.1| ASCL3 [Homo sapiens] >gi|80522...	49	2e-05
gi|20454833|sp|Q9NQ33|ASH3_HUMAN Achaete-scute homolog 3 (b...	49	2e-05
gi|8648972|emb|CAB94840.1| dHAND basic helix-loop-helix tra...	48	2e-05
gi|12054812|emb|CAC20671.1| dHand protein [Mus musculus]	48	2e-05

- List of significant sequence alignments to MASH-1.
- Alignments were scored using the BLOSUM62 scoring matrix.
- The *E* values, shown on the far right, measure the probability of finding a particular alignment by chance in the database being searched.
- As the *E* values approach zero, the statistical significance of the alignment increases.

BLAST Report

- BLAST report ranks hits in order of a measure of statistical significance "**E-value**"
- E-value: **number of hits with the same level of similarity expected by chance**, if there were no true matches in the database (e.g. E-value of 0.01 expected once every 100 searches even when there is no true match in the database)
- E-values for **"relevant" hits are very small ($3x10^{-13}$)**, strongly suggesting that the similarities are not the result of random chance

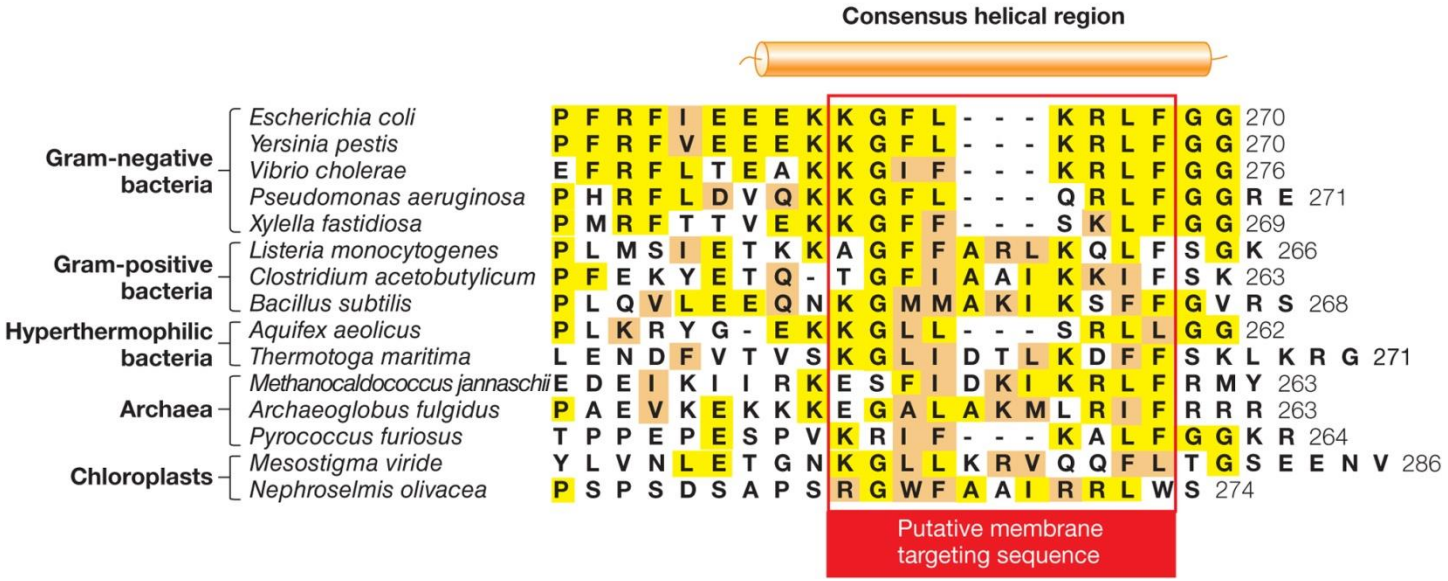
A Pairwise Alignment

- HASH-2, a human homolog of MASH-1
  - Amino acids that are identical in the alignment are shown in the row of characters between the two sequences
  - "**+**" indicates conservative amino acid substitution
  - "**–**" indicates gap/insertion

```
 Score =  105 bits (261), Expect = 2e-22
 Identities = 73/170 (42%), Positives = 92/170 (54%), Gaps = 25/170 (14%)

Query: 85  RQRSSSPELMRCKRRLNFSGFGYSLPQQQPXXXXXXXXXXXXXXKLVNLGFATLREHVPN 144
           R+R +SPEL+RC RR   +    +                      KLVNLGF  LR+HVP+
Sbjct: 23  RRRPASPELLRCSRRRRPAT---AETGGGAAAVARRNERERNRVKLVNLGFQALRQHVPH 79

Query: 145 GAANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAGVLSPTISPN---------- 193
           G A+KK+SKVETLRSAVEYIRALQ+LL EHDAV  A   G+    + P+
Sbjct: 80  GGASKKLSKVETLRSAVEYIRALQRLLAEHDAVRNALAGGLRPQAVRPSAPRGPPGTTPV 139

Query: 194 ----------YSNDLNSMAGSPVSSYSSDE-GSYDPLSPEEQELLDFTNW 232
                     +S  GSP S+YSSD+ G    LSP E+ELLDF++W
Sbjct: 140 AASPSRASSSPGRGGSSEPGSPRSAYSSDDSGCEGALSPAERELLDFSSW 189
```

Bioinformatics

- ORFs are presumed to encode protein (coding sequences (CDS))
- BLAST (*b*asic *l*ocal *a*lignment *s*earch *t*ool) software
  - base by base comparison of 2+ gene sequences
  - assign tentative function of gene or protein structure

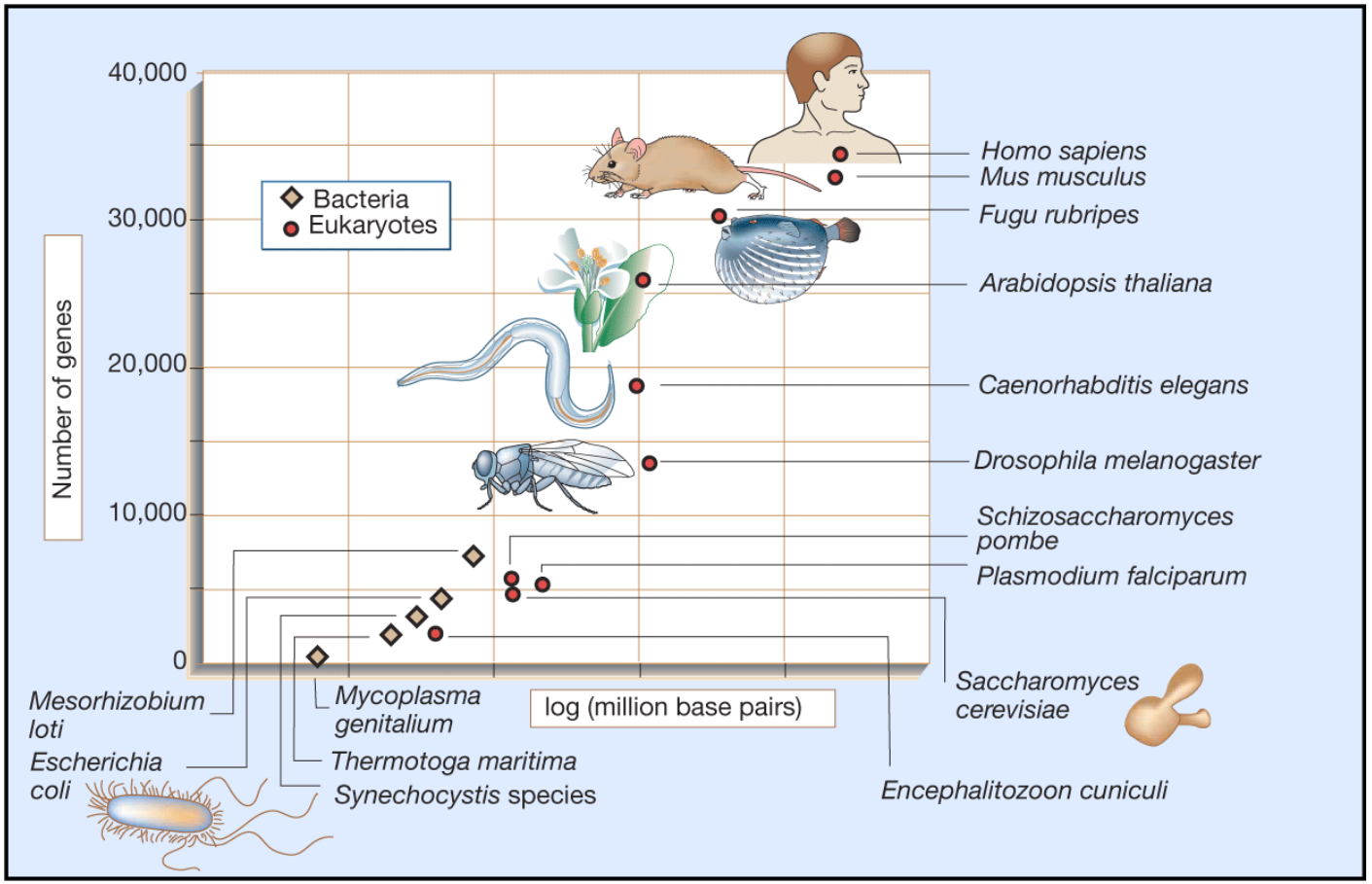


Amino acids residues identical to *E. coli* boxed yellow, and conserved substitutions (e.g. one hydrophobic residue for another) are boxed in organge. Dashes indicate the absence of amino acis in those positions; such gaps may be included to maintain an alignment. (The number of last residue shown relative to the entire amino acid sequence is shown at the end.

Fig. 18.9 Analysis of Conserved Regions of Phylogenetically well-Conserved Proteins. C-terminal amino acid residues of **MinD (cell division protein)** from 15 organims and chloroplasts the conserved region is predicted to form a coil needed for proper localisation to the cell membrane.

Sizes of genomes & numbers of genes

- The smallest prokaryotic genomes are the size of the largest viruses, and the largest prokaryotic genomes have more genes than some eukaryotes.



Relationship between genome size and the number of genes thought to be found in each genome.

Thermotoga maritima

- Metabolic pathways and transport systems (genome analysis)

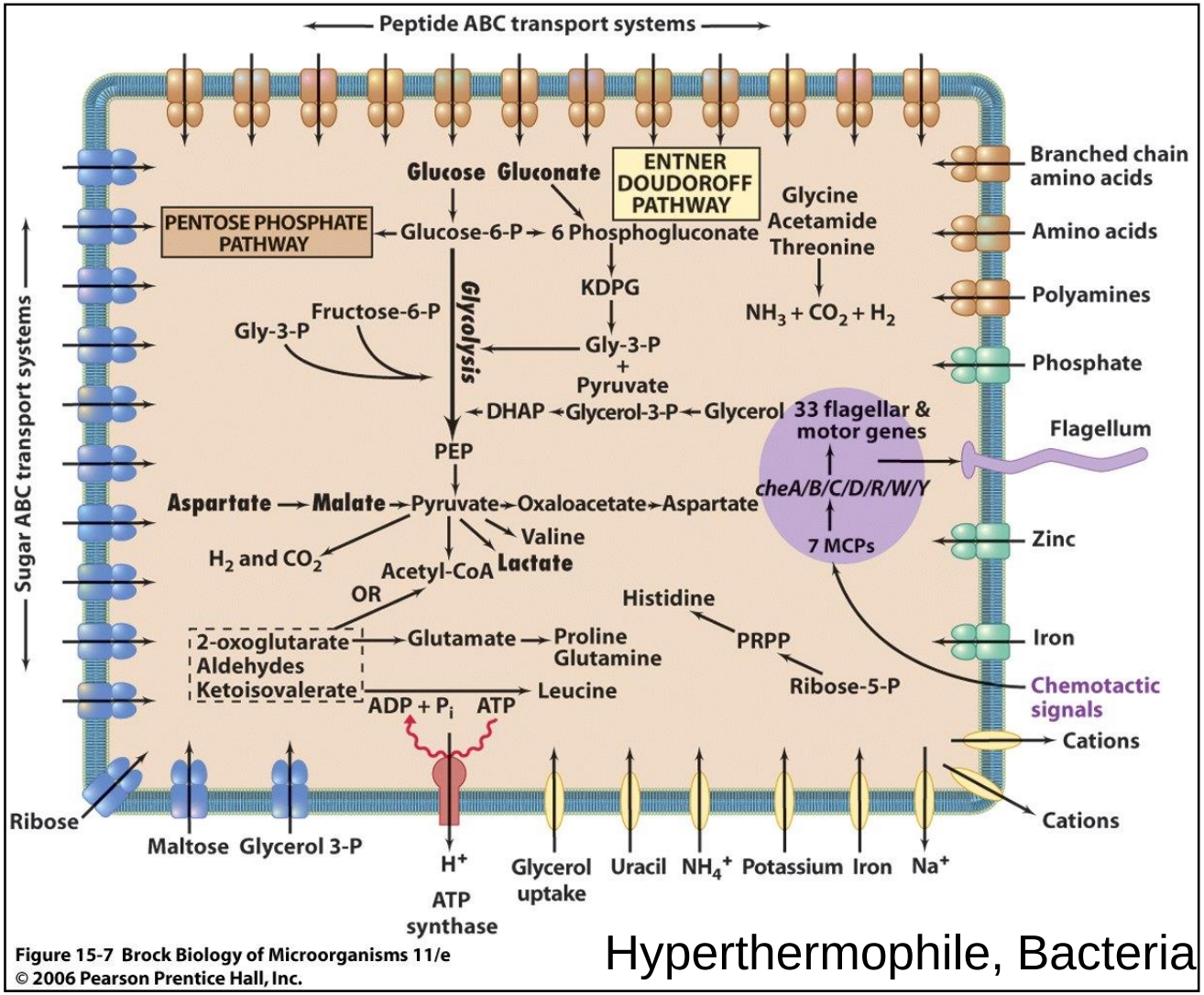


Figure 15-7 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.

Hyperthermophile, Bacteria