JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81
Practical Introduction to Metagenomics

Topic: Introduction to metagenomics

Instructor: Joshua Orvis

Metagenomics

Metagenomics is the culture-independent molecular analysis of environmental samples of cohabiting microbial populations.

– Jo Handelsman (1998)

A *microbiome* signifies the ecological community of commensal, symbiotic, microorganisms that literally share our body space.

– Joshua Lederberg (2001)

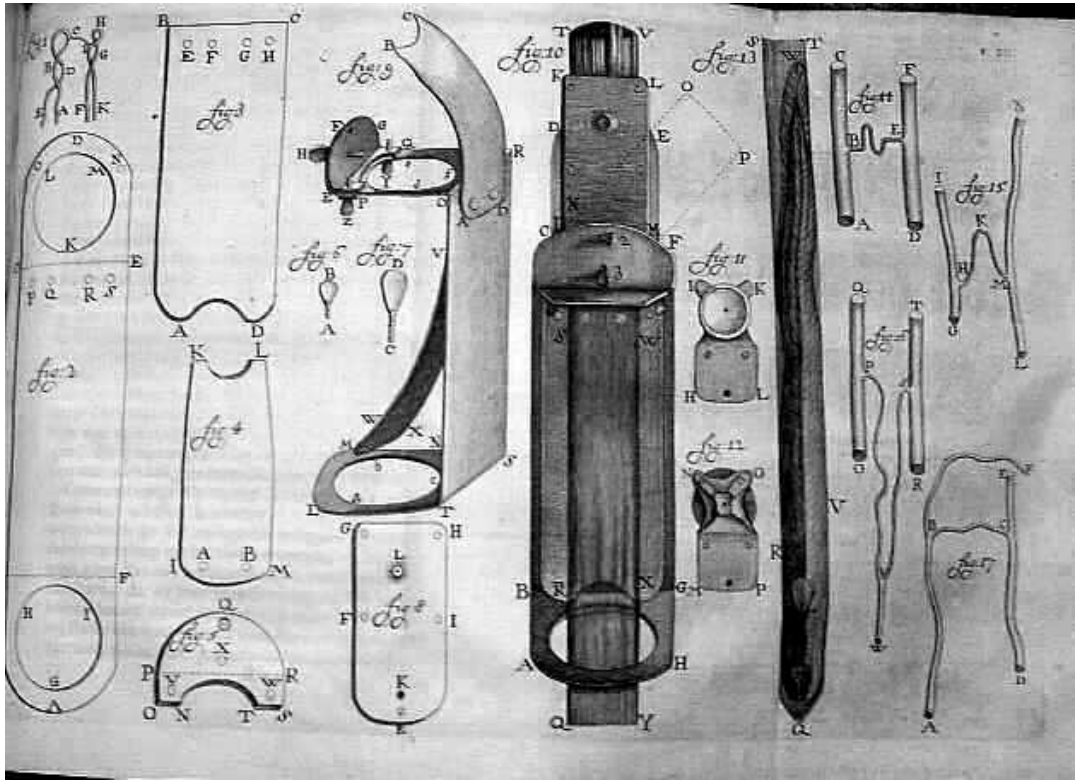
I know you signed up for metagenomics, but first let's step back a bit ….

In the beginning ...

“I then most always saw, with great wonder, that in the said matter there were many very little living animalcules, very prettily a-moving.

Moreover, the other animalcules were in such enormous numbers, that all the water. . . seemed to be alive."

– Leeuwenhoek, 1683, in the first observations of living bacteria ever recorded.



Leeuwenhoek and his 'microscopes'

Discovery over pints

No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

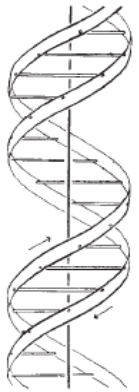
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[3,4] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

"This structure has novel features which are of considerable biological interest."

With this dramatic understatement Watson and Crick began the barely one-page publication which describes the structure of nucleic acids, previously thought to be of too simple composition to be the mechanism of heredity in cells.

With this, our cells could now be viewed as carriers of coded information. Before classified based on morphology and biochemical capabilities, there was now a code to not only uniquely describe each organism, but also to explain how they function and compare to each other.

This would mark the beginning of a great gathering of genetic data.

First free-living organism sequenced

Gradual improvements in both methods and technologies led to our ability to sequence targeted genes of interest and eventually, described here, the entire genome of a free-living organism.

The known sequence space grew rapidly, with each organism chosen generally due to affect on human health or unique biological properties.

The catalogue of prokaryotes grew while sequencers improved, and the many challenges of eukaryotic sequencing became tractable.

RESEARCH ARTICLE

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence. Several viral and organellar genomes have been completely sequenced. Bacteriophage ϕX174 [5386 base pairs (bp)] was the first to be sequenced, by Fred Sanger and colleagues in 1977 (*1*). Sanger *et al.* were also the first to use strategy based on random (unselected) pieces of DNA, completing the genome sequence of bacteriophage λ (48,502 bp) with cloned restriction enzyme fragments (*1*). Subsequently, the 229-kb genome of cytomegalovirus (CMV) (*2*), the 192-kb genome of vaccinia (*3*), and the 187-kb mitochondrial and 121-kb chloroplast genomes of *Marchantia polymorpha* (*4*) have been sequenced. The 186-kb genome of variola (smallpox) was the first to be completely sequenced with automated technology (*5*).

At the present time, there are active genome projects for many organisms, including *Drosophila melanogaster* (*6*), *Escherichia coli* (*7*), *Saccharomyces cerevisiae* (*8*), *Bacillus subtilis* (*9*), *Caenorhabditis elegans* (*10*), and *Homo sapiens* (*11*). These projects, as well as viral genome sequencing, have been based primarily on the sequencing of clones usually derived from extensively mapped restriction fragments, or λ or cosmid clones. Despite advances in DNA sequencing technology (*12*) the sequencing of genomes has not progressed beyond clones on the order of the size of λ (~40 kb). This has been primarily because of the lack of sufficient computational approaches that would enable the efficient assembly of a large number (tens of thousands) of independent, random sequences into a single assembly.

The computational methods developed to create assemblies from hundreds of thousands of 300- to 500-bp complementary DNA (cDNA) sequences (*13*) led us to test the hypothesis that segments of DNA several megabases in size, including entire microbial chromosomes, could be sequenced rapidly, accurately, and cost-effectively by applying a shotgun sequencing strategy to whole genomes. With this strategy, a single random DNA fragment library may be prepared, and the ends of a sufficient number of randomly selected fragments may be sequenced and assembled to produce the complete genome. We chose the free-living organism *Haemophilus influenzae* Rd as a pilot project because its genome size (1.8 Mb) is typical among bacteria, its G+C base composition (38 percent) is close to that of human, and a physical clone map did not exist.

Haemophilus influenzae is a small, nonmotile, Gram-negative bacterium whose only natural host is human. Six *H. influenzae* serotype strains (a through f) have been identified on the basis of immunologically distinct capsular polysaccharide antigens. Non-typeable strains also exist and are distinguished by their lack of detectable capsular polysaccharide. They are commensal residents of the upper respiratory mucosa of children and adults and cause otitis media and respiratory tract infections, mostly in children. More serious invasive infection is caused almost exclusively by type b strains, with meningitis producing neurological sequelae in up to 50 percent of affected children. A vaccine based on the type b capsular antigen is now available and has dramatically reduced the incidence of the disease in Europe and North America.

Genome sequencing. The strategy for a shotgun approach to whole genome sequencing is outlined in Table 1. The theory follows from the Lander and Waterman (*14*) application of the equation for the Poisson distribution. The probability that a base is not sequenced is $P_o = e^{-m}$, where m is the sequence coverage. Thus after 1.83 Mb of sequence has been randomly generated for the *H. influenzae* genome ($m = 1, 1 \times$ coverage), $P_o = e^{-1} = 0.37$ and approximately 37 percent of the genome is unsequenced. Fivefold coverage (approximately 9500 clones sequenced from both insert ends and an average sequence read length of 460 bp) yields $P_o = e^{-5} = 0.0067$, or 0.67 percent unsequenced. If L is genome length and n is the number of random sequence segments done, the total gap length is Le^{-m}, and the average gap size is L/n. Fivefold coverage would leave about 128 gaps averaging about 100 bp in size.

To approximate the random model during actual sequencing, procedures for library construction (*15*) and cloning (*16*) were developed. Genomic DNA from *H. influenzae* Rd strain KW20 (*17*) was mechanically sheared, digested with BAL 31 nuclease to produce blunt ends, and size-fractionated by agarose gel electrophoresis. Mechanical shearing maximizes the randomness of the DNA fragments. Fragments between 1.6 and 2.0 kb in size were excised and recovered. This narrow range was chosen to minimize variation in growth of clones. In addition, we chose this maximum size to minimize the number of complete genes that might be present in a single fragment, and thus might be lost as a result of expression of deleterious gene products. These fragments were ligated to Sma I–cut, phosphatase-treated pUC18 vector, and the ligated products were fractionated on an agarose gel. The linear vector plus insert band was excised and recovered. The ends of the linear recombinant molecules were repaired with T4 polymerase, and the molecules were then ligated into circles. This two-

J.-F. Tomb, B. A. Dougherty, and H. O. Smith are with the Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. J. M. Merrick is with the State University of New York, Department of Microbiology, Buffalo, NY, 14214, USA. K. McKenney is with the National Institute for Standards and Technology, Gaithersburg, MD 20878, USA. All other authors are with The Institute for Genomic Research (TIGR), Gaithersburg, MD, 20878, USA. The address for TIGR as of 9 September 1995 is 9712 Medical Center Drive, Rockville, MD 20850, USA.

*Present address: The National Center for Genome Resources, Santa Fe, NM, 87505, USA.
†To whom correspondence should be addressed.

An exploratory focus

Rather than focus on select organisms, researchers began to consider the microbial makeup of entire environments at a time.

The usual approach would be to culture as many isolates as possible from a sample and then amplify and sequence each one individually.

New methods had to be devised after it was shown that 99% of organisms could not be cultured with standard methods. (PMID:7535888)

Norman Pace and his group were pioneers here, performing PCR amplification and direct sequencing of environmental samples by targeting the highly conserved regions of ribosomal RNA sequences. (PMID:2409920)

(These will be covered in detail later but can be considered as the equivalent of a microbial fingerprint, helping to taxonomically place new sequences.)

Necessity breeds invention

Capillary sequencers (such as the ABI 3730 picture on left) and shotgun methods requiring molecular cloning had served well for single-organism and targeted sequencing projects but didn't scale well for metagenomic goals. The cloning process itself was also an undesired bias.

There has been a flood of “next-gen” high-throughput sequencers which generate orders of magnitude more data and don't require cloning. There are a lot of competitors in this area and we'll cover many of them later in the semester.

A single Illumina run, for example (and pictured left) can produce > 30GB of 100+bp reads.

Data overload ?

You can hardly read a paper in genomics that doesn't begin with a lamentation about the rapid explosion of biological data, much of which is directly attributable to high-throughput sequencing machines.

Processor speed, server-class disk drive capacity, network bandwidth capability, etc. have not kept pace with the scale of DNA sequencing.

Bioinformatics and computational analyses are now the most common bottlenecks in a metagenomics project.

Modern data sets require TB/PB-scale disk arrays, high-speed networks, and distributed grid computing capabilities for most analysis steps.

I want this class to be as representative of real research projects as possible, so you will be operating on data and hardware at this scale. We will use the Data Intensive Academic Grid (DIAG) throughout this course.

Data Intensive Academic Grid (DIAG)

You'll find registration instructions in Blackboard as an Announcement.

Hardware specs:

<u>High-throughput Computational Nodes</u>
- 125 Dual-Processor (Intel Westmere 6-core) nodes
- 48 GB RAM per node
- 3.5 TB local storage per node
- 1500 physical or 3000 hyper-threaded cores

<u>High-performance Computational Nodes</u>
- 5 Quad-Processor (Intel Westmere 8-core) nodes
- 256-512 GB RAM per node
- 1 TB local storage per node
- 160 physical or 320 hyper-threaded cores
- InfiniBand QDR interconnects

<u>Misc.</u>
- SSH access with your own home area
- Ergatis pipeline management tool available

You'll find an quick tutorial on logging in and using DIAG under the 'Resources' tab in Blackboard.

Marching toward the Holy Grail

From strictly visual observation and classification in light microscopes, to decoding an organism's complete genetic content, to initial attempts at exploring variation in communities based on conserved sequences, researchers drove advances in sequencing technology ever-closer to the holy grail of genomic analysis of a given environment (at least the data collection portion of it.)

Environmental genomic analysis holy grail #1:

Culture-free, amplification-free and complete sequence profile of all organisms in any given sample. Cheaply and quickly too.

What does this mean, and have we made it yet?

Next time, let's just sail a boat around the world ...

We can do this without cultures (which would filter what organisms we actually get) and without amplification (which would mess with relative abundance studies.) The rest of these criteria we're still working on.

Why is it so difficult? There is a LOT out there, in almost any environment.

How much is out there?

Prokaryotes: The unseen majority.
Whitman, et al.
Proc. Natl. Acad. Sci. USA. Vol. 95, pp.6578-6583, June 1998. PMID: 9618454

The importance of prokaryotes in most habitats cannot be understated. Here, Whitman, Coleman and Wiebe attempt to reduce the evaluation of total number of prokaryotes on earth, as well as the total amount of organic carbon they represent, to a tractable problem. To do this, they evaluated these primary habitats:

- Aquatic environments – oceans, lakes, sea ice, etc.
- Soil – from forests to deserts.
- Subsurface – terrestrial habitats below 8m and marine sediments below 10cm.
- "Other" - Animals, leaves, insects, air, etc.

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6578–6583, June 1998

Perspective

Prokaryotes: The unseen majority

William B. Whitman†, David C. Coleman‡, and William J. Wiebe§*

*Departments of *Microbiology, ‡Ecology, and §Marine Sciences, University of Georgia, Athens GA 30602*

ABSTRACT **The number of prokaryotes and the total amount of their cellular carbon on earth are estimated to be $4–6 \times 10^{30}$ cells and 350–550 Pg of C (1 Pg = 10^{15} g), respectively. Thus, the total amount of prokaryotic carbon is 60–100% of the estimated total carbon in plants, and inclusion of prokaryotic carbon in global models will almost double estimates of the amount of carbon stored in living organisms. In addition, the earth's prokaryotes contain 85–130 Pg of N and 9–14 Pg of P, or about 10-fold more of these nutrients than do plants, and represent the largest pool of these nutrients in living organisms. Most of the earth's prokaryotes occur in the open ocean, in soil, and in oceanic and terrestrial subsurfaces, where the numbers of cells are 1.2×10^{29}, 2.6×10^{29}, 3.5×10^{30}, and $0.25–2.5 \times 10^{30}$, respectively. The numbers of heterotrophic prokaryotes in the upper 200 m of the open ocean, the ocean below 200 m, and soil are consistent with average turnover times of 6–25 days, 0.8 yr, and 2.5 yr, respectively. Although subject to a great deal of uncertainty, the estimate for the average turnover time of prokaryotes in the subsurface is on the order of $1–2 \times 10^3$ yr. The cellular production rate for all prokaryotes on earth is estimated at 1.7×10^{30} cells/yr and is highest in the open ocean. The large population size and rapid growth of prokaryotes provides an enormous capacity for genetic diversity.**

Although invisible to the naked eye, prokaryotes are an essential component of the earth's biota. They catalyze unique and indispensable transformations in the biogeochemical cycles of the biosphere, produce important components of the earth's atmosphere, and represent a large portion of life's genetic diversity. Although the abundance of prokaryotes has been estimated indirectly (1, 2), the actual number of prokaryotes and the total amount of their cellular carbon on earth have never been directly assessed. Presumably, prokaryotes' very ubiquity has discouraged investigators, because an estimation of the number of prokaryotes would seem to require endless cataloging of numerous habitats.

To estimate the number and total carbon of prokaryotes on earth, several representative habitats were first examined. This analysis indicated that most of the prokaryotes reside in three large habitats: seawater, soil, and the sediment/soil subsurface. Although many other habitats contain dense populations, their numerical contribution to the total number of prokaryotes is small. Thus, evaluating the total number and total carbon of prokaryotes on earth becomes a solvable problem.

Aquatic Environments. Numerous estimates of cell density, volume, and carbon indicate that prokaryotes are ubiquitous in marine and fresh water (e.g., 3–5). Although a large range of cellular densities has been reported (10^4–10^7 cells/ml), the mean values for different aquatic habitats are surprisingly similar. For the continental shelf and the upper 200 m of the open ocean, the cellular density is about 5×10^5 cells/ml. A portion of these cells are the autotrophic marine cyanobacteria and *Prochlorococcus* spp., which have an average cellular density of 4×10^4 cells/ml (6). The deep (>200 m) oceanic water contains 5×10^4 cells/ml on average. From global estimates of volume, the upper 200 m of the ocean contains a total of 3.6×10^{28} cells, of which 2.9×10^{27} cells are autotrophs, whereas ocean water below 200 m contains 6.5×10^{28} cells (Table 1).

The upper 10 cm of sediment in the open ocean is included in the oceanic habitat because, as a result of animal mixing and precipitation, it is essentially contiguous with the overlying water column. Most of the marine sediment is found in the continental rise and abyssal plain, so the numbers of prokaryotes were calculated from an arithmetic average of the cellular densities in the studies cited by Deming and Baross (ref. 9; Table 1). The Nova Scotian continental rise was excluded from this calculation because of its unusual hydrology (10).

There are fewer estimates of the number of prokaryotes in freshwaters and saline lakes (5). Given an average density of 10^6 cells/ml, the total number of cells in freshwaters and saline lakes is 2.3×10^{26}. This value is three orders of magnitude below the numbers of prokaryotes in seawater.

In the polar regions, a relatively dense community of algae and prokaryotes forms at the water–ice interface in annual sea ice (11). In Antarctic sea ice, the estimated number of prokaryotes (2.2×10^{24} cells) was based on the mean cell numbers of Delille and Rosiers (12) and the mean areal extent of seasonal ice (13). If the population size in the Arctic is similar (14), the global estimate for both polar regions is 4×10^{24} cells, only a fraction of the total number of prokaryotes.

Soil. Soil is a major reservoir of organic carbon on earth and an important habitat for prokaryotes. Prokaryotes are an essential component of the soil decomposition subsystem, in which plant and animal residues are degraded into organic matter and nutrients are released into food webs (15). Many studies indicate that the number of prokaryotes in forest soils is much less than the number in other soils. The total number of prokaryotes in forest soil was estimated from detailed direct counts from a coniferous forest ultisol (16), which were considered representative of forest soils in general (Table 2). For other soils, including grasslands and cultivated soils, the numbers of prokaryotes appear about the same, e.g., the number of prokaryotes in Negev desert soil is comparable to the number in cultivated soil (19). Therefore, the numbers of prokaryotes in all other soils were estimated from the unpublished field studies of E. A. Paul for cultivated soils (cited in ref. 18).

Subsurface. The subsurface is defined here as terrestrial habitats below 8 m and marine sediments below 10 cm. Few direct enumerations of subsurface prokaryotes have been made, largely because of the difficulty in obtaining uncontaminated samples. Nevertheless, circumstantial evidence suggests that the subsurface biomass of prokaryotes is enormous (20). For instance, groundwater from deep aquifers and formation

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†To whom reprint requests should be addressed: e-mail: whitman@uga.cc.uga.edu.

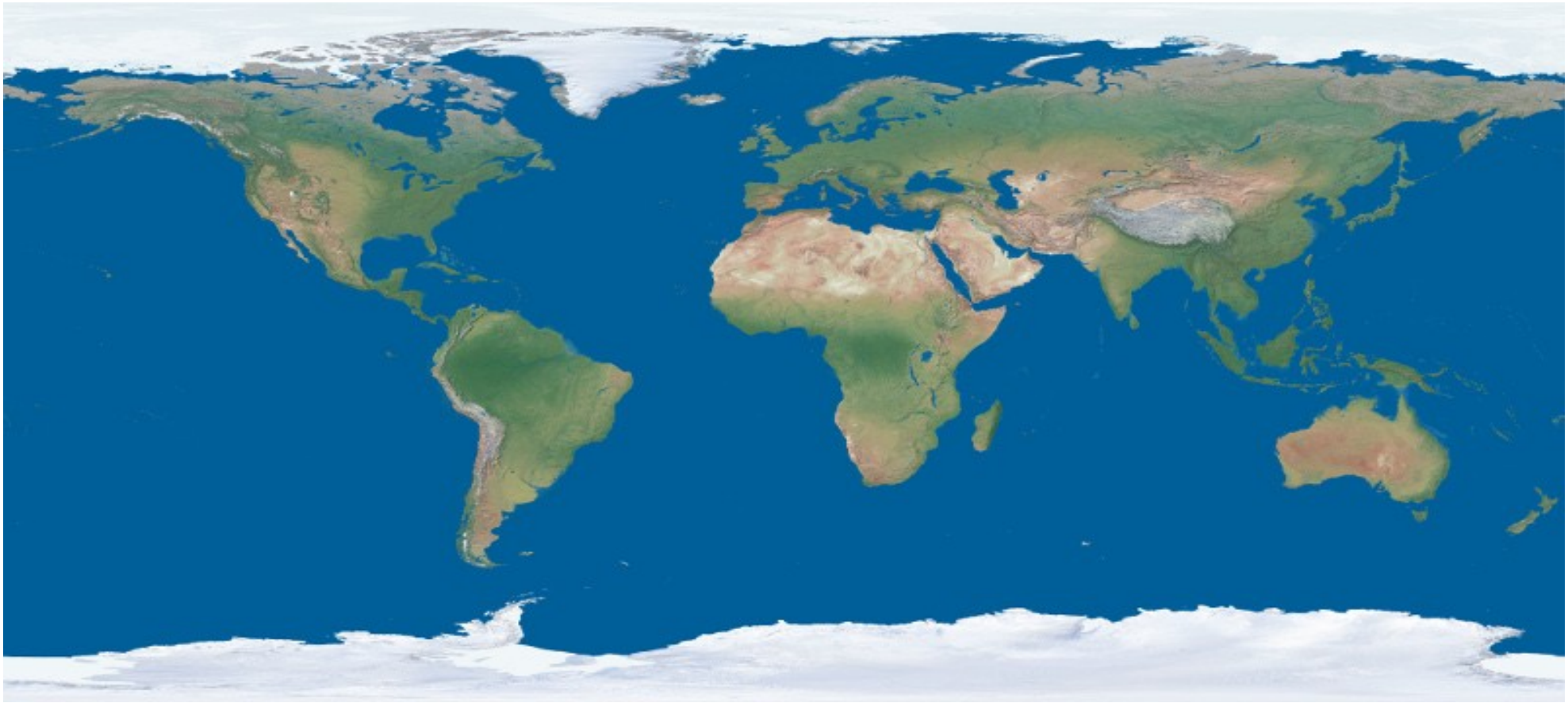
How much is out there?

4-6 x 10^{30} prokaryotic cells

353 – 546 Pg of carbon (C)
1 Pg = 10^{15}g

Production rate = 1.7x10^{30} cells per year

Based on these estimates, the mass of carbon in prokaryotes may be as high as that of all plant matter on Earth combined.



Whitman et al. (a few key points)

“Thus, the total amount of prokaryotic carbon is 60-100% of the estimated total carbon in plants, and inclusion of prokaryotic carbon in global models will almost double estimates of the amount of carbon stored in living organisms."

"Assuming a prokaryotic mutation rate of $4x10^{-7}$ mutations per gene per DNA replication (86, 87), four simultaneous mutations in every gene shared by the populations of marine heterotrophs (in the upper 200m), marine autotrophs, soil prokaryotes, or prokaryotes in domestic animals would be expected to occur once every 0.4, 0.5, 3.4, or 170 hr, respectively."

"Given prokaryotes' numerical abundance and importance in biogeochemical transformations, the absence of detailed knowledge of prokaryotic diversity is a major omission in our knowledge of life on earth."

Fine, so what do they do?

The Whitman paper quantifies that there are a LOT of microbes pretty much everywhere. There are also a lot of grains of sand on the earth, but we care about the microbes because of what they can **do**.

Individual organisms are and will always be interesting and worthy of study, but one approach is to ignore the individual and, instead, see a community of organisms as a pool of shared biochemical functionality.

Microbial Engines

Microbial Ecology

REVIEW

The Microbial Engines That Drive Earth's Biogeochemical Cycles

Paul G. Falkowski,[1]* Tom Fenchel,[2]* Edward F. Delong[3]*

Virtually all nonequilibrium electron transfers on Earth are driven by a set of nanobiological machines composed largely of multimeric protein complexes associated with a small number of prosthetic groups. These machines evolved exclusively in microbes early in our planet's history yet, despite their antiquity, are highly conserved. Hence, although there is enormous genetic diversity in nature, there remains a relatively stable set of core genes coding for the major redox reactions essential for life and biogeochemical cycles. These genes created and coevolved with biogeochemical cycles and were passed from microbe to microbe primarily by horizontal gene transfer. A major challenge in the coming decades is to understand how these machines evolved, how they work, and the processes that control their activity on both molecular and planetary scales.

Earth is ~4.5 billion years old, and during the first half of its evolutionary history, a set of metabolic processes that evolved exclusively in microbes would come to alter the chemical speciation of virtually all elements on the planetary surface. Consequently, our current environment reflects the historically integrated outcomes of microbial experimentation on a tectonically active planet endowed with a thin film of liquid water (*1*). The outcome of these experiments has allowed life to persist even though the planet has been subjected to extraordinary environmental changes, from bolide impacts and global glaciations to massive volcanic outgassing (*2*). Although such perturbations led to major extinctions of plants and animals (*3*), to the best of our knowledge, the core biological machines responsible for planetary biogeochemical cycles have survived intact.

The explosion of microbial genome sequence data and increasingly detailed analyses of the structures of key machines (*4*) has yielded insight into how microbes became the biogeochemical engineers of life on Earth. Nevertheless, a grand challenge in science is to decipher how the ensemble of the core microbially derived machines evolved and how they interact, and the mechanisms regulating their operation and maintenance of elemental cycling on Earth. Here we consider the core set of genes responsible for fluxes of key elements on Earth in the context of a global metabolic pathway.

Essential Geophysical Processes for Life

On Earth, tectonics and atmospheric photochemical processes continuously supply substrates and remove products, thereby creating geochemical cycles (*5*, *6*). These two geophysical processes allow elements and molecules to interact with each other, and chemical bonds to form and break in a cyclical manner. Indeed, unless the creation of bonds forms a cycle, planetary chemistry ultimately will come to thermodynamic equilibrium, which would lead inevitably to a slow depletion of substrates essential for life on the planetary surface. Most of the H_2 in Earth's mantle escaped to space early in Earth's history (*7*); consequently, the overwhelming majority of the abiotic geochemical reactions are based on acid/base chemistry, i.e., transfers of protons without electrons. The chemistry of life, however, is based on redox reactions, i.e., successive transfers of electrons and protons from a relatively limited set of chemical elements (*6*).

The Major Biogeochemical Fluxes Mediated by Life

Six major elements—H, C, N, O, S, and P—constitute the major building blocks for all biological macromolecules (*8*). The biological fluxes of the first five of these elements are driven largely by microbially catalyzed, thermodynamically constrained redox reactions (Fig. 1). These involve two coupled half-cells, leading to a linked system of elemental cycles (*5*). On geological time scales, resupply of C, S, and P is dependent on tectonics, especially volcanism and rock weathering (Fig. 1). Thus, biogeochemical cycles have evolved on a planetary scale to form a set of nested abiotically driven acid-base and biologically driven redox reactions that set lower limits on external energy required to sustain the cycles. These reactions fundamentally altered the surface redox state of the planet. Feedbacks between the evolution of microbial metabolic and geochemical processes create the average redox condition of the oceans and atmosphere. Hence, Earth's redox state is an emergent property of microbial life on a planetary scale. The biological oxidation of Earth is driven by photosynthesis, which is the only known energy transduction process that is not directly dependent on preformed bond energy (*9*).

The fluxes of electrons and protons can be combined with the six major elements to construct a global metabolic map for Earth (Fig. 2). The genes encoding the machinery responsible for the redox chemistry of half-cells form the basis of the major energy-transducing metabolic pathways. The contemporary pathways invariably require multimeric protein complexes (i.e., the microbial "machines") that are often highly conserved at the level of primary or secondary structure. These complexes did not evolve instantaneously, yet the order of their appearance in metabolism and analysis of their evolutionary origins are obscured by lateral gene transfer and extensive selection. These processes make reconstruction of how electron transfer reactions came to be catalyzed extremely challenging (*10*).

In many cases, identical or near-identical pathways may be used for the forward and reverse reactions required to maintain cycles. For example, methane is formed by methanogenic Archaea from the reduction of CO_2 with H_2. If the hydrogen tension is sufficiently low, however, then the reverse process becomes thermodynamically favorable; methane is oxidized anaerobically by Archaea closely related to known, extant methanogens that apparently use co-opted methanogenic machinery in reverse. Low hydrogen tension occurs when there is close spatial association with hydrogen-consuming sulfate reducers (*11*–*13*); thus, this process requires the synergistic cooperation of multispecies assemblages, a phenomenon that is typical for most biogeochemical transformations. Similarly, the citric acid cycle oxidizes acetate stepwise into CO_2 with a net energy yield. In green sulfur bacteria, and in some Archaebacteria, the same cycle is used to assimilate CO_2 into organic matter with net energy expenditure. Indeed, this may have been the original function of that cycle (*14*). Typically, in one direction, the pathway is oxidative, dissimilatory, and produces adenosine 5′-triphosphate, and in the opposite direction, the pathway is reductive, assimilatory, and energy consuming.

However, reversible metabolic pathways in biogeochemical cycles are not necessarily directly related, and sometimes are catalyzed by diverse, multispecies microbial interactions. The various oxidation and reduction reactions that drive Earth's nitrogen cycle (which, before humans, was virtually entirely controlled by microbes) are a good example. N_2 is a highly inert gas, with an atmospheric residence time of ~1 billion years. The only biological process that makes N_2 accessible for the synthesis of proteins and nucleic acids is nitrogen fixation, a reductive process that transforms N_2 to NH_4^+. This biologically irreversible reaction is catalyzed by an extremely conserved heterodimeric enzyme complex, nitrogenase, which is inhibited by oxygen (*15*). In the presence of oxygen, NH_4^+ can be oxidized to nitrate in a two-stage pathway, initially requiring a specific group of Bacteria or Archaea that oxidize ammonia to NO_2^- (via hydroxyamine), which is subsequently oxidized to NO_3^- by a different suite of nitrifying bacteria (*16*). All of the nitrifiers use the small differences in redox potential in the oxidation reactions to reduce CO_2 to

[1]Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences and Department of Earth and Planetary Sciences, Rutgers University, New Brunswick, NJ 08901, USA. [2]Marine Biological Laboratory, University of Copenhagen, Strandpromenaden 5, DK-3000 Helsingør, Denmark. [3]Department of Civil and Environmental Engineering and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: falko@marine.rutgers.edu (P.G.F); tfenchel@bio.ku.dk (T.F.); delong@mit.edu (E.F.D.)

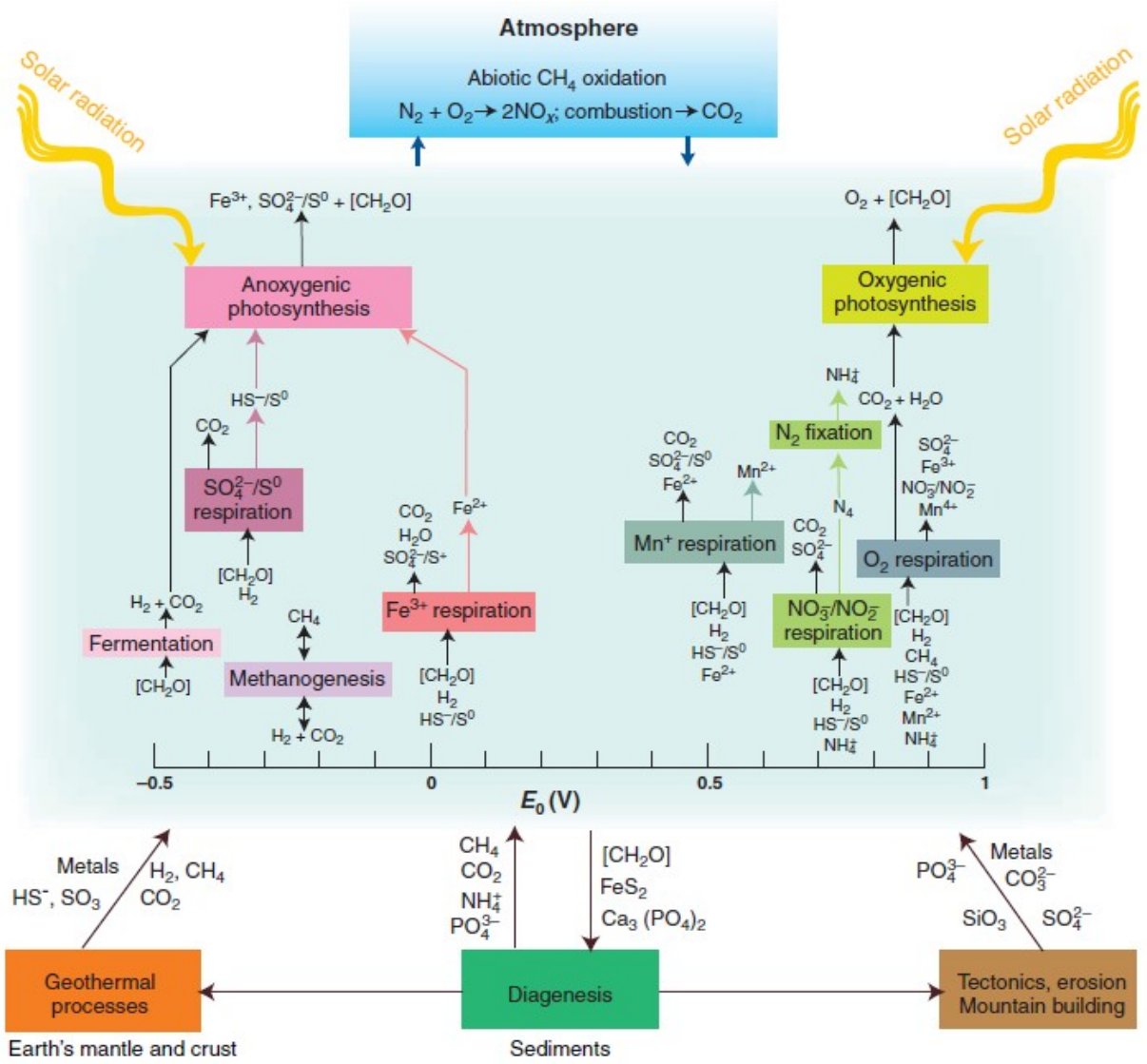
Paul G. Falkowski, Tom Fenchel, Edward F Delong. Science 320, 1034 (2008).
PMID: 18497287

In this review paper the authors consider "microbial engines" and describe that they became the "biochemical engineers of life on earth."

Motivation and scope

“A grand challenge in science is to decipher how the ensemble of the core microbially derived machines evolved and how they interact, and the mechanisms regulating their operation and maintenance of elemental cycling on Earth.”

The authors are interested in examining the flow of key elements on a *global* scale (such as those from Fig. 1 on the right) along with the metabolic pathways they're a part of and core genes that are responsible for them.



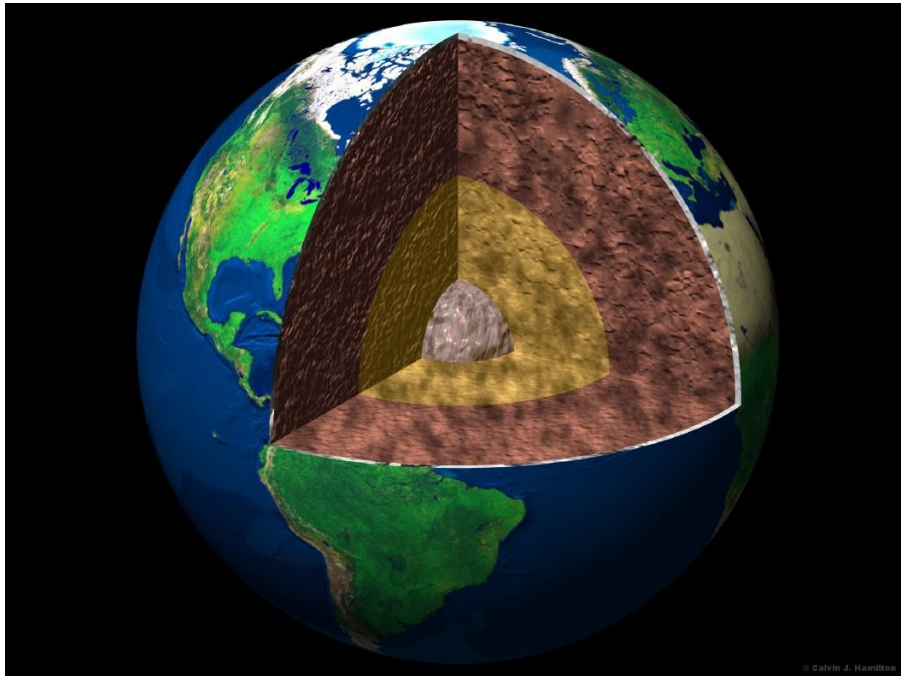
I love this paper

Yes it's overflowing with chemistry, but these guys knew how to tackle something and not be afraid of scale.

They refer to our oceans as a “thin film of liquid water” because, when you consider the entire earth, that makes complete sense.

The sort of thinking exhibited here is also common in many metagenomics studies. The focus is often less on any individual organism than the microbiome as a whole.

How do all the organisms work together or oppose each other, and what is their metabolic and enzymatic activity when considered together?



Most importantly, these data form the base of most of the reporting of a metagenomic environment. The organisms, genes, proteins, etc. described in a metagenomic study ultimately serve as the abstract machines that drive these global processes, and it's important to keep them in mind as we go along.

Further (optional) reading

“Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.”

http://www.ncbi.nlm.nih.gov/pubmed/7542800

“Phylogenetic identification and in situ detection of individual microbial cells without cultivation.”

http://www.ncbi.nlm.nih.gov/pubmed/7535888