JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Survey of major projects

Instructor: Joshua Orvis

Getting here

In the last lesson we introduced the field of metagenomics and the series of advances that led to it. The studies of Whitman et al. showed the sheer and almost unimaginable scale of the presence of microbes on the planet, calling our lack of detailed knowledge of their diversity a major omission. The biogeochemical importance of these microbes was described by Falkowski, Fenchel and Delong.

So these microbes are everywhere and they are critically important, but how are they studied as communities?

The earliest studies here are also (understandably) the least complex and serve as a good starting point to learning about metagenomics principles.

Bacteria underneath a human toenail (Murawski / NGS)

A few things to keep in mind this week.

The three studies we cover this week have a lot of detail regarding the methods and data used. We *will* be going over these and even reproducing parts of it, but I can't introduce all of it at once.

For now, focus on the overall design and approach of each study. It's important to get an overview of the methods as you read through them, with the understanding that we'll get deeper into these in later weeks.

In the acid mine paper, for example, understand that the authors generated a certain number of short sequences as primary data, then computationally assembled and partitioned them into longer sequences. This ultimately provided almost complete assembly of 2 genomes and partially recovered 3 more.

Keep the details like this in your mind that describe the study while at the same time allowing some of the *how* behind it to be explored later.

Acid mine drainage study

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

Tyson GW, Banfield JF, et al.
Nature. 2004 Mar 4;428(6978):37-43.
PMID: 14961025

An abandoned mine in the Sierra Nevada mountains provides an opportunity to do shotgun sequencing directly on a simple population of microorganisms living in an extremely acidic environment.

Specifically, the samples were taken from a low-complexity biofilm growing within a pyrite ore body in an acid mine drainage site.

articles

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

Gene W. Tyson[1], Jarrod Chapman[3,4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul M. Richardson[4], Victor V. Solovyev[4], Edward M. Rubin[4], Daniel S. Rokhsar[3,4] & Jillian F. Banfield[1,2]

[1]Department of Environmental Science, Policy and Management, [2]Department of Earth and Planetary Sciences, and [3]Department of Physics, University of California, Berkeley, California 94720, USA
[4]Joint Genome Institute, Walnut Creek, California 94598, USA

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

The study of microbial evolution and ecology has been revolutionized by DNA sequencing and analysis[1–3]. However, isolates have been the main source of sequence data, and only a small fraction of microorganisms have been cultivated[4–6]. Consequently, focus has shifted towards the analysis of uncultivated microorganisms via cloning of conserved genes[5] and genome fragments directly from the environment[7–9]. To date, only a small fraction of genes have been recovered from individual environments, limiting the analysis of microbial communities as networks characterized by symbioses, competition and partitioning of community-essential roles. Comprehensive genomic data would resolve organism-specific pathways and provide insights into population structure, speciation and evolution. So far, sequencing of whole communities has not been practical because most communities comprise hundreds to thousands of species[10].

Acid mine drainage (AMD) is a worldwide environmental problem that arises largely from microbial activity[11]. Here, we focused on a low-complexity AMD microbial biofilm growing hundreds of feet underground within a pyrite (FeS_2) ore body[12–15]. This represents a self-contained biogeochemical system characterized by tight coupling between microbial iron oxidation and acidification due to pyrite dissolution[11,16,17]. Random shotgun sequencing of DNA from entire microbial communities is one approach for the recovery of the gene complement of uncultivated organisms, and for determining the degree of variability within populations at the genome level. We used random shotgun sequencing of the biofilm to obtain the first reconstruction of multiple genomes directly from a natural sample. The results provide novel insights into community structure, and reveal the strategies that underpin microbial activity in this environment.

Initial characterization of the biofilm

Biofilms growing on the surface of flowing AMD in the five-way region of the Richmond mine at Iron Mountain, California[12], were sampled in March 2000. Screening using group-specific[18] fluorescence *in situ* hybridization (FISH) revealed that all biofilms contained mixtures of bacteria (*Leptospirillum*, *Sulfobacillus* and, in a few cases, *Acidimicrobium*) and archaea (*Ferroplasma* and other members of the Thermoplasmatales). The genome of one of these archaea, *Ferroplasma acidarmanus* fer1, isolated from the Richmond mine, has been sequenced previously (http://www.jgi.doe.gov/JGI_microbial/html/ferroplasma/ferro_homepage.html).

A pink biofilm (Fig. 1a) typical of AMD communities was selected for detailed genomic characterization (see Supplementary Information). The biofilm was dominated by *Leptospirillum* species and contained *F. acidarmanus* at a relatively low abundance (Fig. 1b, c). This biofilm was growing in pH 0.83, 42 °C, 317 mM Fe, 14 mM Zn, 4 mM Cu and 2 mM As solution, and was collected from a surface area of approximately 0.05 m^2.

A 16S ribosomal RNA gene clone library was constructed from DNA extracted from the pink biofilm, and 384 clones were end-sequenced (see Supplementary Information). Results indicated the presence of three bacterial and three archaeal lineages. The most abundant clones are close relatives of *L. ferriphilum*[19] and belong to *Leptospirillum* group II (ref. 13). Although 94% of the *Leptospirillum* group II clones were identical, 17 minor variants were detected with up to 1.2% 16S rRNA gene-sequence divergence from the dominant type. Tightly defined groups (up to 1% sequence divergence) related to *Leptospirillum* group III (ref. 13), *Sulfobacillus*, *Ferroplasma* (some identical to fer1), 'A-plasma'[15] and 'G-plasma'[15] were also detected. *Leptospirillum* group III, G-plasma and A-plasma have only recently been detected in culture-independent molecular surveys. FISH-based quantification (Fig. 1c; see also Supplementary Information) confirmed the dominance of *Leptospirillum* group II in the biofilm.

Community genome sequencing and assembly

In conventional shotgun sequencing projects of microbial isolates, all shotgun fragments are derived from clones of the same genome. When using the shotgun sequencing approach on genomes from an

What is acid mine/rock drainage?

Any rock/ores containing sulfide minerals with metal ions like iron, copper, or nickel are exposed to water they will oxidize and generate acidity in the surrounding area. In small amounts, this can be washed away easily.

But human activities that expose a lot of rock at once, such as mining, can cause this to happen in a very large scale. When mines are abandoned they fill with water after the drainage pumps stop, and acid rock drainage is the usual result.

Tulsaquah Chief Mine, BC. (photo credit Chris Miller)

Why an acid mine site?

Since mixed populations of organisms are sampled and sequenced together, any given sequence read could be from any of thousands of different species and strains.

We'll cover this in a later lesson, but computationally partitioning these reads and assembling them into longer contigs (or even complete genomes) is challenging.

The extremely harsh conditions of these habitats limit the organisms living there to acidophiles, whose presence actually accelerate oxidation by several orders of magnitude.

The challenging environment limits the variety of species who live there, making it possible to do pooled sampling, sequencing and analysis.

The Richmond Mine

Tyson et al. sampled from a thin biofilm within a pyrite ore body from this northern California mine. The most acidic water naturally found on Earth was sampled here and reported by the US Geological Survey in 2000, with a pH of -3.6.

Sample notes

- Surface area of the biofilm collected was 0.05m^2
- pH: 0.83
- Temperature: 42 degrees C
- Metals present: 317 mM Fe, 14 mM Zn, 4mM Cu, and 2mM As

Partitioning the sequences by organism

103,462 Sanger reads (avg 737bp each)

↓

Assembled into 1,183 “scaffolds” using JAZZ

↓

Scaffolds separated by average G+C content

↓

Further subdivided by read depth (coverage)

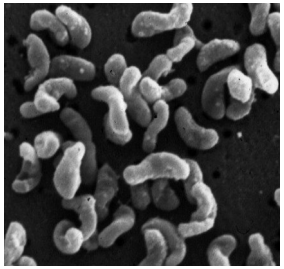
The authors here needed to separate their assembled scaffolds by organism, but how? First, they relied on the knowledge that genomes from different organisms tend to have generalized differences in GC content.

They then further separated the scaffolds in each GC bin by relative abundance. This was done based on the assumptions that the read depth was an approximation of the abundance of each organism and there was no sampling bias.

Organism profile

High G+C @ 10x coverage = 70 scaffolds (2.23 Mb)

Leptospirillum group II – identified by the presence of a single 16S rRNA gene. Further evidence includes 'matching' the estimated genome length of *L. ferrooxidans* and having a matching G+C content with *L. ferriphilum*.



Low G+C @ 10x coverage = 59 scaffolds (1.82 Mb)

Ferroplasma type II – A new species! The 16S rRNA found here matched a fer1 isolate with 99% identity, but the rest of the genome was 22% divergent in sequence **yet** had conserved genome size, local gene order and content. This is strong evidence that one of the dominant species there was previously unknown.

High G+C @ 3x coverage = 474 scaffolds (2.66 Mb)

Leptospirillum group III – Identified by rRNA markers but with significant sequence divergence and local gene order from *Leptospirillum* group II. This includes any scaffolds from *Sulfobacillus*.

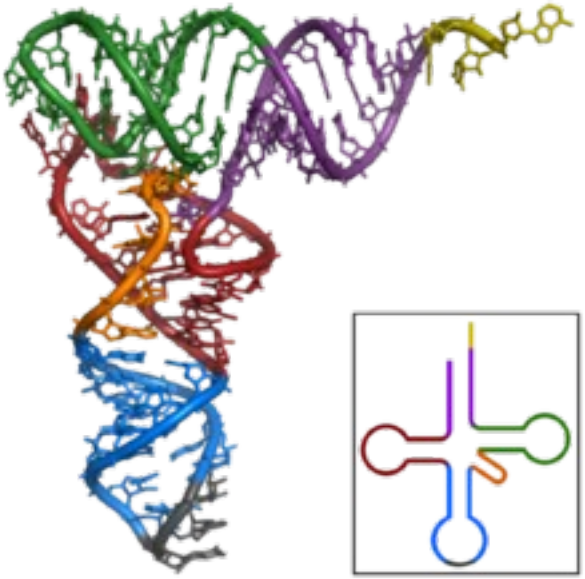
Low G+C @ 3x coverage = 580 scaffolds (4.12 Mb)

Ferroplasma type I – all scaffolds that aligned to the fer1 isolate with >= 96% identity. The rest are G-plasma and, to a much lesser extent, A-plasma.

Method validation

Each partition is formed by at least dozens of scaffolds, so how do the authors claim near-complete genomes or validate their methods of partitioning by G+C and depth of coverage?

The primary evidence is that each partition has a full set (or nearly so) of tRNA synthetases and only one set of rRNA genes.



They assert that these results, along with the agreement between the recovered and anticipated genome sizes, validate this method (at least for this particular microbial community.)

What do you think? This is a good thing to raise in the discussions.

Other findings

An extremely low incidence of polymorphisms was found in the *Leptospirillum* group II scaffolds, possibly indicating strong recent environmental selection or founder effect.

The Ferroplasma type II strains seem to have undergone homologous recombination in a population that seems to be “dominated by strains with mosaic genomes constructed by recombination of three closely related but distinct genome types.”

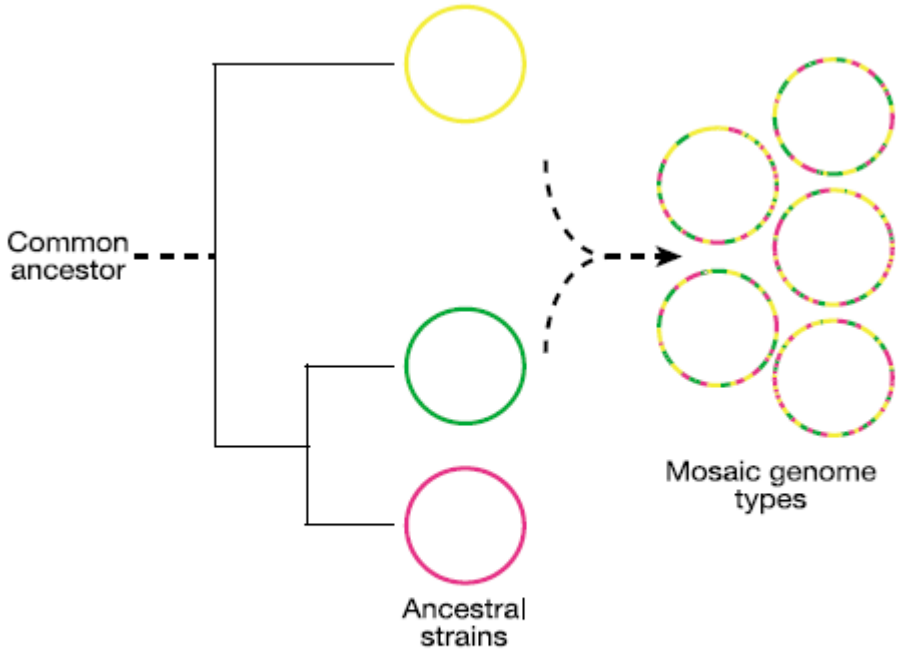


Figure 3 Schematic diagram illustrating a diversity of mosaic genome types within the *Ferroplasma* type II population that are inferred to have arisen by homologous recombination between three closely related ancestral genome types (pink, yellow and green).

Genomes are dynamic

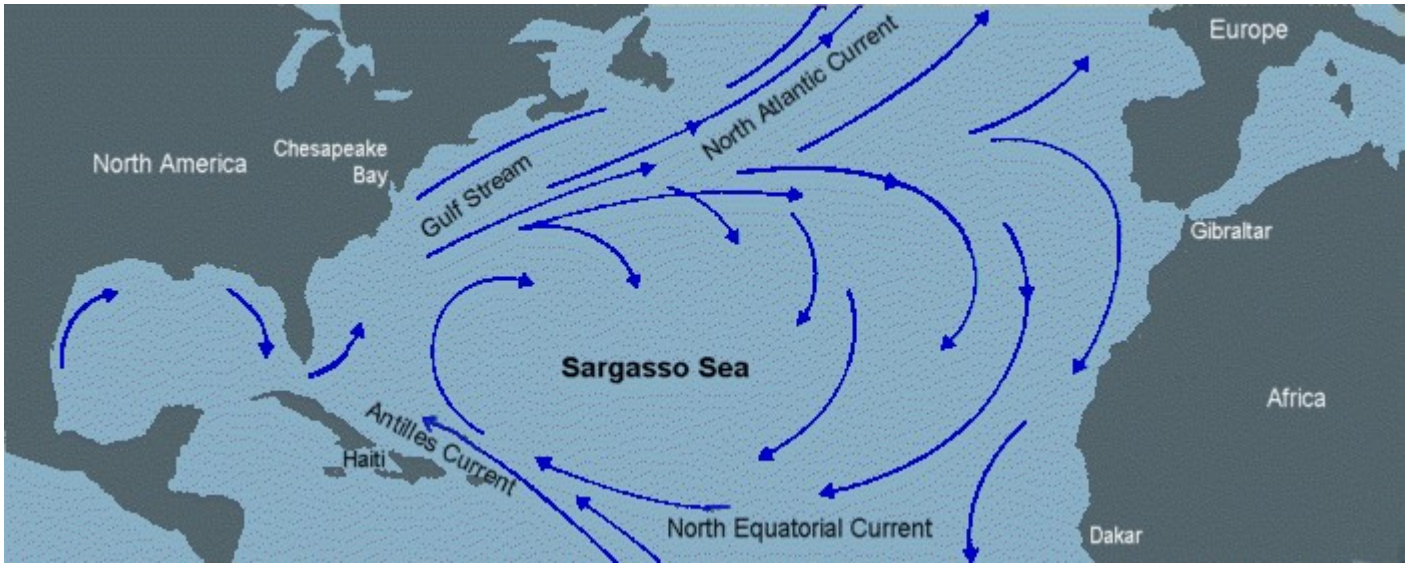
"Genomes are dynamic biological structures: any notion that they are singular, unchanging entities does not capture the process that shapes the diversity we see today. It is only by peering directly into naturally occurring genomic diversity, as Tyson et al. have done, that the tempo, mode and mechanism of genome evolution and diversification, and its relationship to higher-order biological and ecological processes, will become clear."

– Edward F. DeLong, commenting on the acid mine drainage study in a Nature "news and views."

Scaling up

It's tempting at this point to just start grabbing samples from anywhere and sequencing everything found there. But in 2004 sequencing is still relatively expensive and analysis methods are still being worked out, so the initial large-scale studies were chosen very carefully.

Next we cover a study involving water samples taken from the Sargasso Sea off the coast of Bermuda.



Sargasso Sea study

The authors report:

- 1.045 billion non-redundant bases sequenced
- 1800 species estimated
- 148 previously unknown bacterial phylotypes
- 1.2 million previously unknown genes
- 782 new rhodopsin-like photoreceptors

RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1]* Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

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.

Microorganisms are responsible for most of the biogeochemical cycles that shape the environment of Earth and its oceans. Yet, these organisms are the least well understood on Earth, as the ability to study and understand the metabolic potential of microorganisms has been hampered by the inability to generate pure cultures. Recent studies have begun to explore environmental bacteria in a culture-independent manner by isolating DNA from environmental samples and transforming it into large insert clones. For example, a previously unknown light-driven proton pump, proteorhodopsin, was discovered within a bacterial artificial chromosome (BAC) from the genome of a SAR86 ribotype (*1*), and soil microbial DNA libraries have been constructed and screened for specific activities (*2*).

Here we have applied whole-genome shotgun sequencing to environmental-pooled DNA samples to test whether new genomic approaches can be effectively applied to gene and species discovery and to overall environmental characterization. To help ensure a tractable pilot study, we sampled in the Sargasso Sea, a nutrient-limited, open ocean environment. Further, we concentrated on the genetic material captured on filters sized to isolate primarily microbial inhabitants of the environment, leaving detailed analysis of dissolved DNA and viral particles on one end of the size spectrum and eukaryotic inhabitants on the other, for subsequent studies.

The Sargasso Sea. The northwest Sargasso Sea, at the Bermuda Atlantic Time-series Study site (BATS), is one of the best-studied and arguably most well-characterized regions of the global ocean. The Gulf Stream represents the western and northern boundaries of this region and provides a strong physical boundary, separating the low nutrient, oligotrophic open ocean from the more nutrient-rich waters of the U.S. continental shelf. The Sargasso Sea has been intensively studied as part of the 50-year time series of ocean physics and biogeochemistry (*3*, *4*) and provides an opportunity for interpretation of environmental genomic data in an oceanographic context. In this region, formation of subtropical mode water occurs each winter as the passage of cold fronts across the region erodes the seasonal thermocline and causes convective mixing, resulting in mixed layers of 150 to 300 m depth. The introduction of nutrient-rich deep water, following the breakdown of seasonal thermoclines into the brightly lit surface waters, leads to the blooming of single cell phytoplankton, including two cyanobacteria species, *Synechococcus* and *Prochlorococcus*, that numerically dominate the photosynthetic biomass in the Sargasso Sea.

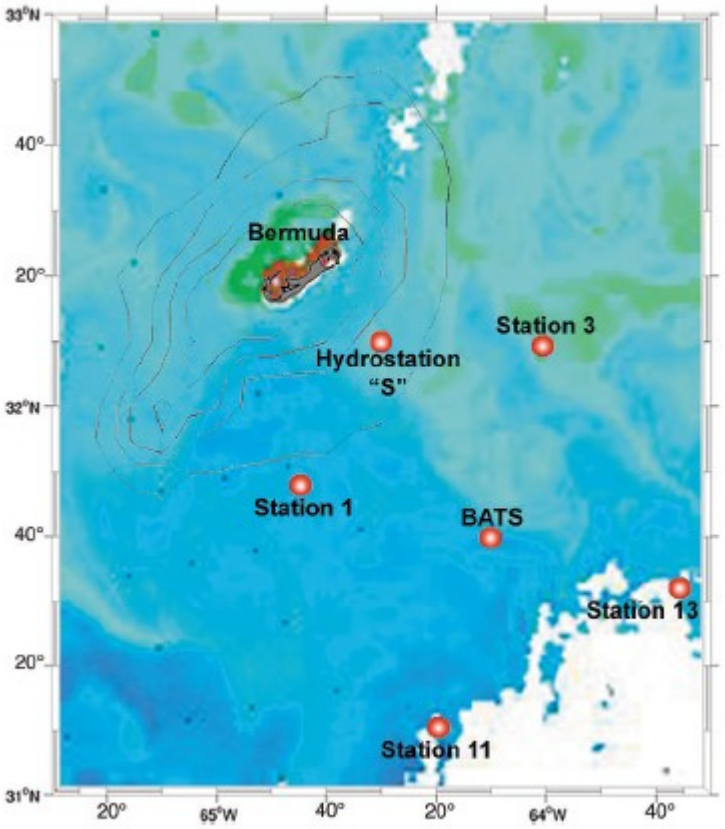
Surface water samples (170 to 200 liters) were collected aboard the RV Weatherbird II from three sites off the coast of Bermuda in February 2003. Additional samples were collected aboard the SV Sorcerer II from "Hydrostation S" in May 2003. Sample site locations are indicated on Fig. 1 and described in table S1; sampling protocols were fine-tuned from one expedition to the next (*5*). Genomic DNA was extracted from filters of 0.1 to 3.0 μm, and genomic libraries with insert sizes ranging from 2 to 6 kb were made as described (*5*). The prepared plasmid clones were sequenced from both ends to provide paired-end reads at the J. Craig Venter Science Foundation Joint Technology Center on ABI 3730XL DNA sequencers (Applied Biosystems, Foster City, CA). Whole-genome random shotgun sequencing of the Weatherbird II samples (table S1, samples 1 to 4) produced 1.66 million reads averaging 818 bp in length, for a total of approximately 1.36 Gbp of microbial DNA sequence. An additional 325,561 sequences were generated from the Sorcerer II samples (table S1, samples 5 to 7), yielding approximately 265 Mbp of DNA sequence.

Environmental genome shotgun assembly. Whole-genome shotgun sequencing projects have traditionally been applied to identify the genome sequence(s) from one particular organism, whereas the approach taken here is intended to capture representative sequence from many diverse organisms simultaneously. Variation in genome size and relative abundance determines the depth of coverage of any particular organism in the sample at a given level of sequencing and has strong implications for both the application of assembly algorithms and for the metrics used in evaluating the resulting assembly. Although we would expect abundant species to be deeply covered and well assembled, species of lower abundance may be represented by only a few sequences. For a single genome analysis, assembly coverage depth in unique regions should approximate a Poisson distribution. The mean of this distribution can be estimated from the observed data, looking at the depth of coverage of contigs generated before any scaffolding. The assembler used in this study, the Celera Assembler (*6*), uses this value to heuristically identify clearly unique regions to form the backbone of the final assembly within the scaffolding phase. However, when the starting material consists of a mixture of genomes of varying abundance, a threshold estimated in this way would classify samples from the most abundant organism(s) as repetitive, due to their greater-than-average depth of coverage, paradoxically leaving the most abundant organisms poorly assembled. We therefore used manual curation of an initial

[1]The Institute for Biological Energy Alternatives, [2]The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA. [3]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [4]The J. Craig Venter Science Foundation Joint Technology Center, 5 Research Place, Rockville, MD 20850, USA. [5]University of Southern California, 223 Science Hall, Los Angeles, CA 90089–0740, USA. [6]Bermuda Biological Station for Research, Inc., 17 Biological Lane, St George GE 01, Bermuda.

*To whom correspondence should be addressed. E-mail: jcventer@tcag.org

Why the Sargasso Sea?



Part of the Bermuda Atlantic Time-series Study site (BATS), this northwest part of the Sargasso Sea is one of the most studied and best characterized sites in the open ocean.

This study complements the already-underway 50-year study of ocean physics and biogeochemistry there.

It's a nutrient-limited, open-ocean environment.

Its photosynthetic biomass is dominated by single-cell phytoplankton, including two cyanobacteria species.

(and Craig likes taking his boat out)

Sampling

RV Weatherbird II

Sampled 170-200 liters from 3 sites in Feb. 2003

1.66 million reads for 1.36 Gbp of sequence

SV Sorcerer II

Sampled Hydro-station S in May 2003

325 thousand reads for 265 Mbp of sequence data

Assembly and binning

The reads were assembled into contigs (and ultimately scaffolds) using the Celera Assembler. Starting with the well-sampled material (with at least 3x coverage), analysis was performed on 333 scaffolds made up of 2226 contigs and spanning 30.9 Mbp.

These resulting scaffolds were then binned according to the following criteria:

- Depth of coverage
- Oligo-nucleotide frequencies
- Similarity to previously sequenced genomes

The first of these criteria we've seen before. The second references a study (PMID:7482779) that found that the set of dinucleotide odds ratios for an organism are very stable and that they constitute a signature of each DNA genome.

The third introduces the idea of using a set of reference genomes to classify a metagenomic sample. The rapid rate of both isolate-sequencing and other metagenomic studies where uncultured organisms are sequenced to a high depth of coverage generates an ever-increasing pool of reference data to use for classification and annotation. We'll explore this in a later lecture.

Between two text-heavy slides

from: reddit.com

Organism profile

More than half of the assembled scaffolds were assigned to a selection of abundant species. They are:

- *Burkholderia* relative
- 2 distinct strains related to the *Shewanella oneidensis*
- *SAR86*
- A conglomerate of *Prochlorococcus* strains
- An uncultured marine archaeon
- 10 putative “mega plasmids”

Again, having the large amount of sequences available allows for evolutionary analysis via inspection of polymorphisms, insertions, deletions and re-arrangement events. Here that is seen with the *Prochlorococcus* strains, which illustrate a population continuum rather than a distinct species.

Other results

The rest of the paper describes aspects of analysis we'll cover in greater detail as the semester progresses, including functional annotation, more detailed phylogenetic diversity estimates, etc.

For now it's important to read over the study with the goal of getting an overview of its design and analysis goals, allowing for the **how** of many steps to be returned to in the coming weeks.

Global Ocean Survey

The Sargasso study can be thought of as a pilot study for this one, in which Venter and collaborators sailed around the oceans, now sampling from different locations at various depths.

(disclaimer: I worked on this study)

The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1], Kenneth Nealson[1,12], Robert Friedman[1], Marvin Frazier[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Department of Biological Sciences, University of Southern California, Avalon, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America, **4** Your Genome, Your World, Rockville, Maryland, United States of America, **5** Departmento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, **6** Department of Oceanography, University of Hawaii, Honolulu, Hawaii, United States of America, **7** Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada, **8** Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama, **9** Departamento de Oceanografía, Universidad de Concepción, Concepción, Chile, **10** Escuela de Química, Universidad de Costa Rica, San Pedro, Costa Rica, **11** Department of Environmental Sciences, Rutgers University, New Brunswick, New Jersey, United States of America, **12** Department of Earth Sciences, University of Southern California, Los Angles, California, United States of America

The world's oceans contain a complex mixture of micro-organisms that are for the most part, uncharacterized both genetically and biochemically. We report here a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analyzed as part of the *Sorcerer II* Global Ocean Sampling expedition. These samples, collected across a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp). Though a few major microbial clades dominate the planktonic marine niche, the dataset contains great diversity with 85% of the assembled sequence and 57% of the unassembled data being unique at a 98% sequence identity cutoff. Using the metadata associated with each sample and sequencing library, we developed new comparative genomic and assembly methods. One comparative genomic method, termed "fragment recruitment," addressed questions of genome structure, evolution, and taxonomic or phylogenetic diversity, as well as the biochemical diversity of genes and gene families. A second method, termed "extreme assembly," made possible the assembly and reconstruction of large segments of abundant but clearly nonclonal organisms. Within all abundant populations analyzed, we found extensive intra-ribotype diversity in several forms: (1) extensive sequence variation within orthologous regions throughout a given genome; despite coverage of individual ribotypes approaching 500-fold, most individual sequencing reads are unique; (2) numerous changes in gene content some with direct adaptive implications; and (3) hypervariable genomic islands that are too variable to assemble. The intra-ribotype diversity is organized into genetically isolated populations that have overlapping but independent distributions, implying distinct environmental preference. We present novel methods for measuring the genomic similarity between metagenomic samples and show how they may be grouped into several community types. Specific functional adaptations can be identified both within individual ribotypes and across the entire community, including proteorhodopsin spectral tuning and the presence or absence of the phosphate-binding gene *PstS*.

Citation: Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol 5(3): e77. doi:10.1371/journal.pbio.0050077

This article is part of the Oceanic Metagenomics collection.

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

Received July 14, 2006; **Accepted** January 16, 2007; **Published** March 13, 2007

Abbreviations: CAMERA, Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis; GOS, Global Ocean Sampling; NCBI, National Center for Biotechnology Information

* To whom correspondence should be addressed. E-mail: DRusch@venterinstitute.org

This article is part of Global Ocean Sampling collection in *PLoS Biology*. The full collection is available online at http://collections.plos.org/plosbiology/gos-2007.php.

GOS paper and more

As you read through the GOS paper consider the similarities to the Sargasso one and the various ways the focus has been expanded. We will expand on this in the discussion boards.

This paper, along with the Human Microbiome Project papers published in the Summer of 2012, will form the basis of much of the practical portion of the course.

The practical shift starts next week, when we cover sequencing and assembly.

Further (optional) reading

"Fluorescence in situ hybridization: past, present and future"

http://jcs.biologists.org/content/116/14/2833.full - A review of the applications of FISH in the period of this acid mine paper, which have expanded even since then.